HiER 2015

Elbeshausen, Faaß, Griesbaum, Heuwing und Jürgens (Hrsg)

HiER 2015

Proceedings des 9. Hildesheimer
Evaluierungs- und Retrievalworkshop

Universitätsverlag Hildesheim

Stefanie Elbeshausen, Gertrud Faaß, Joachim Griesbaum, Ben Heuwing, Julia Jürgens (Hrsg.): HIER 2015 - Proceedings des 9. Hildesheimer Evaluierungs- und Retrievalworkshop, Hildesheim 2015

Universitätsverlag Hildesheim
Universitätsplatz 1
31141 Hildesheim
verlag@uni-hildesheim.de
ISBN (Open Access) 978-3-934105-59-1
ISBN-A (Open Access) 10.978.3934105/591
Hildesheim 2015

Veranstalter & Tagungsteam

Universität Hildesheim
Institut für Informationswissenschaft und Sprachtechnologie
Universitätsplatz 1
31141 Hildesheim
http://www.uni-hildesheim.de/iwist.htm

Inhaltsverzeichnis

Vorwort

Die Digitalisierung formt unsere Informationsumwelten. Disruptive Technologien dringen verstärkt und immer schneller in unseren Alltag ein und verändern unser Informations- und Kommunikationsverhalten. Informationsmärkte wandeln sich.

Der 9. Hildesheimer Evaluierungs- und Retrievalworkshop HIER 2015 thematisiert die Gestaltung und Evaluierung von Informationssystemen und die Ausprägung von Informationsmärkten vor dem Hintergrund der sich beschleunigenden Digitalisierung. Im Fokus stehen die folgenden Themen:

- Digital Humanities
- Internetsuche und Online Marketing
- Information Seeking und nutzerzentrierte Entwicklung
- E-Learning und Informationsmarkt

Dieser Band fasst die Vorträge des 9. Hildesheimer Evaluierungs- und Retrieval-Workshops (HIER) zusammen, der am 9. und 10. Juli 2015 an der Universität Hildesheim stattfand. Die HIER Workshop-Reihe begann im Jahr 2001 mit dem Ziel, die Forschungsergebnisse der Hildesheimer Informationswissenschaft zu präsentieren und zu diskutieren. Mittlerweile nehmen immer wieder Kooperationspartner von anderen Institutionen teil, was wir sehr begrüßen. HIER schafft auch ein Forum für Systemvorstellungen und praxisorientierte Beiträge.

Hildesheim, im Juli 2015

Stefanie Elbeshausen, Gertrud Faaß, Joachim Griesbaum,
Ben Heuwing und Julia Jürgens

Session 1
Digital Humanities

Projekt Welt der Kinder

Überblick über die informationswissenschaftliche Bedarfsanalyse

Ben Heuwing, Thomas Mandl, Christa Womser-Hacker

Universität Hildesheim, Universitätsplatz 1, 31141 Hildesheim
[name]@uni-hildesheim.de

Zusammenfassung

In dem Projekt „Welt der Kinder“ erarbeiten Geschichtswissenschaftler zusammen mit Computerlinguisten und Informationswissenschaftlern innovative Methoden, um ein großes historisches Schulbuchkorpus mit Mitteln der automatischen Textanalyse vergleichend auszuwerten. Aus einer informationswissenschaftlichen Perspektive werden dabei Methoden für die Definition von Anforderungen an Analyseprozesse und -Werkzeuge untersucht, die durch komplexe, domänenspezifische Fragestellungen entstehen.

Abstract

For the project “Welt der Kinder” (children of the world), researcher from the area of language technology and information science cooperate with historians to develop innovative methods for the comparative analysis of a large corpus of historic text books using means of automatic text analysis. From an information science perspective, the project intends to investigate methods for the analysis of requirements regarding complex and domain specific information problems.

1. Projektziele

Das Projekt „Welt der Kinder“[1] strebt die Analyse von Deutungen und Sichtweisen auf die Welt aus der Sicht von Kindern im deutschen Kaiserreich (1871-1918) an. Dafür wird von der Annahme ausgegangen, dass das in der Schule vermittelte Wissen das Weltbild von Kindern in dieser Periode aufgrund der geringen Zugangsmöglichkeiten zu anderen Medien in besonderem Maße geprägt hat. Die Untersuchung basiert auf einem Korpus von Ge-

[1] Angesiedelt am Georg-Eckert-Institut, Braunschweig, gemeinsam mit der TU Darmstadt gefördert durch die Leibniz-Gemeinschaft unter SAW-2014-GEl-2

schichtsschulbüchern (Strötgen 2014) mit geplant 3000 Bänden,[2] welches im weiteren Projektverlauf noch um Kinder- und Jungendliteratur ergänzt wird. Aus fachwissenschaftlicher Sicht ordnet sich das Projekt damit in das Gebiet der Kultur- und Wissensgeschichte ein. Der Fokus liegt auf der Auswertung von Geschichtsschulbüchern vor dem Hintergrund der parallelen Nationalisierungs- und Globalisierungsprozesse des 19. Jahrhunderts. Dafür sollen Veränderungen im Zeitverlauf, aber auch Unterschiede etwa zwischen Schulbüchern für bestimmte Regionen oder konfessionelle Ausrichtungen und zwischen Schulbüchern und Kinder- und Jugendliteratur untersucht werden.

Während in der Informationswissenschaft insbesondere Studien zu den Informationsbedürfnissen von Geisteswissenschaftlern, aber auch von Geschichtswissenschaftlern bei der Suche nach Quellen und Sekundärliteratur vorliegen (Allen & Sieczkiewicz 2010; Rhee 2012; Toms & O'Brien 2008), gibt es wenige Untersuchungen zu den Anforderungen an die Unterstützung während des Analyse- und Auswertungsprozesses. In dem Projekt „Welt der Kinder" werden aus diesem Grund Methoden erprobt, welche die iterative und kooperative Definition von Anforderungen ermöglichen, den Austausch zwischen den beteiligten Geschichtswissenschaftlern, Computerlinguisten und Informationswissenschaftlern verbessern und sich in ein allgemeines, interdisziplinäres Vorgehensmodell übertragen lassen. Gleichzeitig werden an diesem Fallbeispiel domänenspezifische Verhaltensweisen bei der Informationsanalyse untersucht.

2. Automatische Textanalyse in der digitalen Geschichtswissenschaft

In den Disziplinen der digitalen Geisteswissenschaften existieren unterschiedliche Herangehensweisen für die Analyse großer Textmengen aus den Bereichen Text-Mining und Computerlinguistik. Trotzdem steht die hermeneutische Auswertung und Interpretation der Texte häufig im Vordergrund des Erkenntnisinteresses und spielt damit eine wichtige Rolle im gesamten Analyseprozess.

Für die digitale Geschichtswissenschaft unterscheidet Robertson (2014) in einem Blogbeitrag zwei allgemeine Ansätze der Textanalyse: Die Identifikation und Verknüpfung von einzelnen semantischen Einheiten in den Texten und die Bestimmung allgemeiner Trends über das gesamte Korpus. Die Identifizierung von semantischen Einheiten wird insbesondere für die Identifizie-

[2] Zugänglich über http://gei-digital.gei.de

rung von Ortsnennungen eingesetzt mit dem Ziel einer Kartierung der Inhalte. Ein umfangreiches Beispiel ist die Analyse von subjektiven räumlichen Wahrnehmungen und der Konstruktion von Räumen anhand der Berichterstattung US-amerikanischer Lokalzeitungen zum Ende des 19. Jahrhunderts (Blevins 2014).

Die Analyse von sprachlichen Mustern über ein gesamtes Korpus wird in der digitalen Geschichtswissenschaft wegen des begrenzten Zugangs zu analysierbaren Textmaterialien im Vergleich zur Literaturwissenschaft seltener durchgeführt (Robertson 2014). Zudem hat sich in der Disziplin noch keine Übereinstimmung über die Interpretierbarkeit und Aussagekraft von mit statistischen Modellen erhobenen Ergebnissen ausgebildet. Auf der Basis von korpusanalytischen Verfahren, insbesondere von Kollokationen und Konkordanzen, können etwa historische Diskurse in Zeitungen vergleichend analysiert werden (Baker et al. 2008). Während die Analyse von Termkollokationen in einem zu untersuchenden Unterkorpus statistisch belegbare Unterscheidungen ermöglicht, bieten Konkordanzen die Möglichkeit zur Analyse von Unterschieden in der Begriffsverwendung im Kontext ihres Auftretens und damit eine Anbindung an qualitative Analyseschritte. Einfache Topic-Modellierung mittels Latent Dirichlet Allocation (LDA, Blei et al. 2003) und Abwandlungen zur Modellierung weiterer latenter Variablen, etwa Veröffentlichungszeitraum (Blei & Lafferty 2006) oder Autoren (Rosen-Zvi et al. 2010), bilden in den Digital Humanities inzwischen ein vieldiskutiertes Standardverfahren, insbesondere für die Literaturwissenschaft, aber etwa auch für komplexe Analysen in der Politikwissenschaft im Projekt e-Identity (Kliche et al. 2014). Newman & Block (2006) oder Nelson (2010) zeigen das Potential von Topic-Modellierung für die Geschichtswissenschaft am Beispiel der Analyse historischer Zeitungskorpora. Die Modellierung beruht auf dem gemeinsamen Auftreten von Termen in definierten Dokumenteinheiten und versucht dabei, gleichzeitig die Verteilung von Termen auf Themenfeldern (Topics), und die Zuordnung von Dokumenten zu diesen Themen ausgehend von vorgegebenen Wahrscheinlichkeitsverteilungen zu optimieren. Als Ergebnis dieses statistischen Optimierungsprozesses stehen die Zuordnungen von allen Termen zu allen Topics und von allen Topics zu allen Dokumenten über Wahrscheinlichkeitswerte. Die Wahl der Parameter für die Modellierung umfasst primär die gewünschte Anzahl von Topics, aber auch Parameter für die Festlegung der Wahrscheinlichkeitsverteilungen, von denen der Optimierungsprozess ausgeht, die Bestimmung von Termen (Auswahl von Wortformen, Anwendung von Stopwortlisten, Schreibweisennormalisierung) und den zu analysierenden Dokumenteneinheiten (etwa Satz, Absatz, Seite, Kapitel, Gesamtdokument). Daher ergibt sich für die Erstellung von Modellen im Allgemeinen ein iterativer Prozess, in welchem Modelle gezielt für eine Forschungsfrage optimiert werden (Blei 2012; DiMaggio et al. 2013).

Für die Analyse von Ergebnissen aus der automatischen Textanalyse werden häufig mehrere Modellierungsebenen kombiniert und die Ergebnisse zusätzlich visualisiert. Daraus können Probleme für die Nachvollziehbarkeit und das Vertrauen in die Belastbarkeit der Ergebnisse entstehen (Chuang et al. 2012). Für das Projekt „Welt der Kinder“ bedeutet dies, dass neben der Interaktion für die Informationssuche und -analyse auch die Modellierung auf die Anwender ausgerichtet werden muss. Für die nutzerzentrierte Untersuchung ergibt sich jedoch erschwerend, dass sich in der Fachdisziplin etablierten Methoden nicht direkt auf quantitative Untersuchungen im Projekt übertragen lassen. Daher wird für die Anforderungserhebung in diesem Projekt eine Kombination von Methoden zur Erhebung existierender Arbeitsweisen (Kontextanalyse) mit Methoden der kooperativen Gestaltung und Entwicklung angestrebt.

3. Nutzerzentrierte Bedarfsanalyse

Experten und Nutzer aus der Fachdisziplin spielen in dem Projekt „Welt der Kinder“ eine zentrale Rolle, um durch die nutzerzentrierte Bedarfsanalyse deren spezifisches Vorgehen bei der Textanalyse und die damit verbundenen Ziele zu unterstützen. Für die Anforderungserhebung erfolgte zunächst eine Kontextanalyse von Vorgehensweisen bei der Inhaltsanalyse von Bildungsmedien aus der Sicht der Geschichtswissenschaft, in einem Fall vergleichend mit stärker sozialwissenschaftlich geprägten Methoden aus der Politikwissenschaft. Die Studie wurde mit sechs Teilnehmern als Interview an deren Arbeitsplatz durchgeführt. Dadurch konnten die verwendeten Werkzeuge, Materialien und konkreten Handlungsschritte berücksichtigt werden, um so den Prozess der iterativen Analyse und Wissensgenerierung nachzuvollziehen („Sense Making“ - Pirolli & Card 2005). Die Ergebnisse dieser Vorstudie lieferten im weiteren Verlauf des Projektes die Grundlage für die Diskussion des vorgesehenen Analyseprozesses. Das Vorgehen stützt sich auf das Verfahren des Contextual Design (Holtzblatt & Beyer 2013).

Die Ergebnisse zeigten insgesamt vergleichbare Schritte bei der Analyse, mit großen Abweichungen in der konkreten Umsetzung. Zunächst erfolgt in allen Fällen die Konstruktion eines auf die Fragestellung ausgerichteten Unterkorpus von Medien, wobei die Fragestellung selbst häufig wieder an die zur Verfügung stehenden Quellen angepasst wird. Dafür kann im Vorfeld eine zusätzliche, quantitativ ausgerichtete Auswertung vorhandener Quellen anhand von Typus und Inhalten durchgeführt werden. Auswertungen können zunächst erst an einzelnen Beispielen durchgeführt werden, um die zu analysierenden Phänomene zu definieren. Im weiteren Verlauf zeigt sich dann jedoch häufig, dass diese Definition weiter vereinfacht werden muss, um für die Analyse mit vertretbarem Aufwand manuell mit weiteren Quellen angewendet werden zu können. Besonders wichtig für die Analyse ist die Kon-

textualisierung, etwa zeitlich zu historischen Ereignissen oder hinsichtlich der Autoren und Sammlungsgeschichte einer Quelle. Die Analyse wird meist manuell durchgeführt, wobei Zwischenergebnisse in Dokumenten gespeichert werden – nach Fragestellungen in Textdokumenten oder nach unterschiedlichen Kriterien strukturiert etwa in Tabellenkalkulationsdokumenten. Genauere Differenzierungen in den einzelnen Medien werden dann anhand einer tiefergehenden, qualitativen Analyse einzelner Beispiele durchgeführt.

Ein weiteres wichtiges Element der Bedarfsanalyse stellt die Evaluierung erster Ergebnisse aus der textstatistischen Analyse des Gesamtkorpus mit Topic-Modellen dar. Dafür wurden mit unterschiedlichen Parametern mehrere Generationen von Topic-Modellen erarbeitet (LDA, s.o.). In einer qualitativ ausgerichteten Evaluierung bewerteten die beteiligten Geschichtswissenschaftler als Experten die Ergebnisse hinsichtlich der Relevanz für die untersuchten Forschungsfragen und die Verständlichkeit der einzelnen Topics. Dabei stellte sich primär heraus, dass die Topics aus der Perspektive der beteiligten Geschichtswissenschaftler eine möglichst hohe Kohärenz aufweisen müssen, um als zielführend für die Fragestellungen des Projektes eingeordnet zu werden, und, dass für die Interpretation der Topics primär die Zuordnungen der Dokumente genutzt werden. Auf Basis dieser Erkenntnisse können automatische Methoden für die Evaluierung der semantischen Kohärenz eingesetzt werden, welche auf externen Bewertungen für die semantische Ähnlichkeit zwischen den Termen eines Topics beruhen (etwa Newman et al. 2010). Weitere Methoden, welche die Interpretierbarkeit von Topics und von Zuordnungen von Dokumenten zu Topics aus der Sicht von Juroren bewerten (Chang et al. 2009), sind aufgrund der projektbezogenen Bewertungsmaßstäbe und den hohen fachlichen Anforderungen, die an Juroren gestellt werden, schwer zu übertragen. Geplant ist trotzdem der Aufbau einer Sammlung von Dokumenten mit Relevanzbewertungen als Goldstandard für den Vergleich von Topic-Modellen hinsichtlich ihres Nutzens für das Dokumentenretrieval.

Abbildung 1: Screenshot - Prototyp für die Analyse

Um die Eignung der Modelle zu untersuchen, wurde weiterhin ein Prototyp für die interpretative Analyse umgesetzt (Abbildung 1). Dieser ermöglicht die Suche über Terme, die Einschränkung von Ergebnissen nach Topics und die Visualisierung von Topic-Verteilungen in Treffermengen über Zeit oder ihre Ausgabe in Form von Tabellen für Vergleiche in anderen Metadatenkategorien wie Schulform oder Erscheinungsort. Zusätzliche Experimente umfassen die Erstellung von *Dynamic Topic Models* (Blei & Lafferty 2006) und die Visualisierung der Veränderungen in der Rangfolge der ersten 100 Terme. Weiterhin wurden auch externe Werkzeuge untersucht und eingesetzt, um Untermengen des Korpus zu analysieren und die Ergebnisse zu visualisieren.

Alle bereits vorhandenen Mittel für die Analyse werden derzeit angewendet, um in interdisziplinärer Zusammenarbeit die Untersuchung bestehender Thesen zu ermöglichen. Dieser Prozess der externen Validierung von analytischen Modellen in der Textanalyse (DiMaggio et al. 2013) bietet zudem Möglichkeiten für die Beteiligten in der Projektgruppe, andere Perspektiven auf die Fragestellungen kennenzulernen und das weitere Vorgehen auf die Belange der Fachdisziplin auszurichten. Dabei wird deutlich, dass die Definition von Anfragen und die Interpretation der Ergebnisse durch die Fachwissenschaftler sowohl für die Überprüfung der Plausibilität der Ergebnisse, als auch für die Einordnung der Granularität und Relevanz hinsichtlich der zu untersuchenden Forschungsfragen von besonderer Wichtigkeit ist. Da bei der externen Validierung existierende Hypothesen untersucht werden (konfirmatorische Analysen), besteht eine weitere Herausforderung darin, in offenen, explorativen Analysen relevante Unterschiede zu differenzieren und hinsichtlich ihrer statistischen Signifikanz zu überprüfen.

Die Ergebnisse der Anforderungsanalyse werden kontinuierlich anhand von Anwendungsszenarien dokumentiert und als statische Prototypen (Wireframes) für die Weiterentwicklung im Projekt umgesetzt. Das Ziel besteht in der Bereitstellung eine Sammlung von existierenden und selbst entwickelten Werkzeugen, welche die Untersuchung der Fragestellungen des Projektes aus unterschiedlichen Perspektiven ermöglicht.

4. Zusammenfassung

Die Ergebnisse der bisherigen Anforderungsanalyse weisen auf die Notwendigkeit einer Analyseplattform hin, welche zuverlässig iterative Untersuchungen von Zusammenhängen in vielen Dimensionen und die transparente Darstellung von Analyseergebnissen ermöglicht. Chancen liegen dabei in der Integration von quantitativen und qualitativen Analysemöglichkeiten. Dabei ist noch unklar, ob Analysemodelle für die Textanalyse allgemein für das gesamte Korpus erstellt werden können, oder ob sie spezifisch für jede Fra-

gestellung erarbeitet werden sollten. Weitere Schritte im Projekt bilden die stärkere Untersuchung von direkten und indirekten Wertungen in Bezug auf die thematisierten Inhalten und die bessere Unterstützung eines explorativen Analyseverhaltens.

5. Literaturverzeichnis

Allen, R. B. ; Sieczkiewicz, R. (2010): How historians use historical newspapers. In: *Proceedings of the American Society for Information Science and Technology* Bd. 47, Nr. 1, S. 1–4

Baker, P. ; Gabrielatos, C. ; KhosraviNik, M. ; Krzyzanowski, M. ; McEnery, T. ; Wodak, R. (2008): A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. In: *Discourse & Society* Bd. 19, Nr. 3, S. 273–306

Blei, D. M. (2012): Topic modeling and digital humanities. In: *Journal of Digital Humanities* Bd. 2, Nr. 1

Blei, D. M. ; Lafferty, J. D. (2006): Dynamic topic models. In: *Proceedings of the 23rd international conference on Machine learning*. ACM, S. 113–120

Blei, D. M. ; Ng, A. Y. ; Jordan, M. I. (2003): Latent dirichlet allocation. In: *the Journal of machine Learning research* Bd. 3, S. 993–1022

Blevins, C. (2014): Space, nation, and the triumph of region: A view of the world from Houston. In: *Journal of American History* Bd. 101, Nr. 1, S. 122–147

Chang, J. ; Gerrish, S. ; Wang, C. ; Boyd-Graber, J. L. ; Blei, D. M. (2009): Reading tea leaves: How humans interpret topic models. In: *Advances in neural information processing systems*, S. 288–296

Chuang, J. ; Ramage, D. ; Manning, C. ; Heer, J. (2012): Interpretation and trust: Designing model-driven visualizations for text analysis. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12*. New York, NY, USA: ACM, S. 443–452

DiMaggio, P. ; Nag, M. ; Blei, D. (2013): Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. In: *Poetics* Bd. 41, Nr. 6, S. 570–606

Holtzblatt, K. ; Beyer, H. R. (2013): Contextual Design. In: *The Encyclopedia of Human-Computer Interaction, 2nd Ed.*

Kliche, F. ; Blessing, Andre ; Heid, U. ; Sonntag, J. (2014): The eIdentity text exploration workbench. In: Calzolari, N. ; Choukri, K. ; Declerck, T. ; u.a. (Hg.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA)

Nelson, R. K. (2010): *Mining the dispatch*. abgerufen 11.06.2015 URL: http://dsl.richmond.edu/dispatch/pages/home

Newman, D. J. ; Block, S. (2006): Probabilistic topic decomposition of an eighteenth-century American newspaper. In: *Journal of the American Society for Information Science and Technology* Bd. 57, Nr. 6, S. 753–767

Newman, D. ; Lau, J. H. ; Grieser, K. ; Baldwin, T. (2010): Automatic evaluation of topic coherence. In: *Human Language Technologies 2010.* ACL, S. 100–108

Pirolli, P. ; Card, S. (2005): The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In: *Proceedings of International Conference on Intelligence Analysis*. Bd. 5. Mitre McLean, VA, S. 2–4

Rhee, H. L. (2012): Modelling historians' information-seeking behaviour with an interdisciplinary and comparative approach. In: *Information Research* Bd. 17, Nr. 4

Robertson, S. (2014): The Differences between digital history and digital humanities. abgerufen 02.02.2015
URL: http://drstephenrobertson.com/blog-post/the-differences-between-digital-history-and-digital-humanities/

Rosen-Zvi, M. ; Chemudugunta, C. ; Griffiths, T. ; Smyth, P. ; Steyvers, M. (2010): Learning Author-topic Models from Text Corpora. In: *ACM Trans. Inf. Syst.* Bd. 28, Nr. 1, S. 4:1–4:38

Strötgen, R. (2014): New information infrastructures for textbook research at the Georg Eckert Institute. In: *History of Education & Children's Literature* Bd. 9, Nr. 1

Toms, E. G. ; O'Brien, H. L. (2008): Understanding the information and communication technology needs of the e-humanist. In: *Journal of Documentation* Bd. 64, Nr. 1, S. 102–130

Erstellung themenspezifischer Korpora mit dem LDA-basierten Klassifikationsmodell

Melanie Dick

Universität Hildesheim, Universitätsplatz 1, 31141 Hildesheim
melaniedick@gmx.net

Zusammenfassung

Das deutsche eIdentity-Zeitungskorpus soll mittels Klassifikation von irrelevanten Artikeln bereinigt werden. In einem zweistufigen Klassifikationsmodell werden zunächst Merkmale extrahiert. Diese werden in einem zweiten Schritt für die Klassifikation einzelner Zeitungsartikel als relevant beziehungsweise irrelevant verwendet. Für die Merkmalsextraktion wird neben dem gängigen Bag of Words-Modell (BoW) die Latent Dirichlet Allocation (LDA) eingesetzt, evaluiert und optimiert.

Abstract

The German newspaper corpus from the eIdentity-project needs to be filtered from non-relevant articles using classification. This paper introduces a two-staged classification-model. In the first stage all features will be extracted from the corpora and in the second stage they will be used for the classification of each article as relevant or non-relevant. The feature extraction will be either done with a common Bag of Words-model (BoW) or with the Latent Dirichlet Allocation (LDA). Both extraction methods will be evaluated and optimized.

1. Einleitung

Das eIdentity-Projekt, gefördert vom Bundesministerium für Bildung und Forschung (BMBF, FK 01UG1234) beschäftigt sich mit dem Problem, wie große Datenmengen zum Beispiel in Form eines Zeitungskorpus, auf (semi-)automatischem Wege für die Politikwissenschaft erschlossen werden können. Für die effiziente Beantwortung politikwissenschaftlicher Fragestellungen, welche sich zum Beispiel auf Ereignisketten über mehrere Jahre hinweg beziehen, werden automatisierte Verfahren benötigt. Ressourcenerschließung

durch den Einsatz computerlinguistischer Methoden soll dabei helfen (vgl. Universität Stuttgart, 2014).

2. Fragestellung und Motivation

Das eIdentity-Projekt setzt sich aus zwei Teilen zusammen. Zum einen sollen „*Multiple kollektive Identitäten in internationalen Debatten um Krieg und Frieden seit dem Ende des Kalten Krieges*“ (Universität Stuttgart 2014) untersucht werden und zum anderen sollen hierfür „*Sprachtechnologische Werkzeuge und Methoden für die Analyse mehrsprachiger Textmengen in den Sozialwissenschaften*“ (ebd.) entwickelt und verwendet werden.

Dieser Artikel beschreibt die Bereinigung des deutschen eIdentity-Korpus mit einem zweistufigen Klassifikationsmodell von für die politikwissenschaftliche Aufgabenstellung nicht relevanten Artikeln. Relevante Artikel sind solche die militärische Interventionen seit Ende des Kalten Kriegs betreffen und zu deren Motivation auf Konzepte von nationaler, europäischer, religiöser usw. Identität verweisen. Bei nicht relevanten Artikeln handelt es sich beispielsweise um Buch- oder Filmkritiken. Im ersten Schritt werden Merkmale aus den Zeitungsartikeln extrahiert. Hierfür wird ein Bag of Words-Modell (BoW) als Baseline und alternativ die Latent Dirichlet Allocation (LDA, vgl. Blei, Ng und Jordan 2003), ein generatives Wahrscheinlichkeitsmodell, verwendet. Der Klassifikator im zweiten Schritt verwendet die Merkmale um die Artikel als relevant beziehungsweise irrelevant zu bewerten. Für die Testläufe wurde das Werkzeug *mallet* (vgl. McCallum 2002) verwendet.

Die Auswertung von manuell annotierten Zeitungsartikeln zeigte bereits, dass die komplexe geisteswissenschaftliche Aufgabenstellung des eIdentity-Projekts Möglichkeiten der subjektiven Auslegung bietet. Die *Interrater-Reliabilität* zwischen den verschiedenen Annotatoren der Zeitungsartikel erreicht mit einem Kappa von 0,615 zwar eine substantiell große Übereinstimmung (vgl. Landis und Koch 1977, S.165), zeigt aber auch, dass nicht alle Artikel eindeutig annotiert wurden. Allerdings bezieht sich das berechnete Kappa nur auf die Artikel, welche mit mehr als einer Annotation vorlagen. Dies entspricht etwa knapp fünf Prozent der manuellen Annotationen (4,9% von insgesamt 1.021 Annotationen). Der Kappa-Wert verdeutlicht, dass die Relevanz einzelner Artikel trotz Fachwissen[3] und weitestgehender Standardisierung der Klassifikationskriterien über einen Leitfaden, vom einzelnen

3 Die Annotation erfolgte durch Studierende der Sozialwissenschaften

Annotator je unterschiedlich bewertet werden kann und auch die manuelle Einzelbewertung keine vollständige Genauigkeit erreicht. Diese lag für die manuellen Annotationen bei 0,84.

3. Das zweistufige Klassifikationsmodell

Das zweistufige Klassifikationsmodell wird in zwei Varianten verwendet, welche sich im ersten Schritt, der Merkmalsextraktion unterscheiden. Die BoW-basierte Variante bildet die Baseline und ist ein gängiges Verfahren zur Merkmalsextraktion für Klassifikationsaufgaben (vgl. Zhang und Zhou 2010, S. 43). Wörter sowie deren Häufigkeit werden beim BoW-Modell als Merkmal des jeweiligen Artikels extrahiert und in einem Vektor gespeichert, welcher an den Klassifikator übergeben wird. Das als Alternative eingesetzte LDA verwendet sogenannte *topics*. Ein Artikel wird von mehreren latenten *topics* repräsentiert. Die Anzahl der *topics* muss allerdings manuell bestimmt werden.

Im zweiten Schritt klassifiziert ein Klassifikationsalgorithmus aufbauend auf den extrahierten Merkmalen die Zeitungsartikel. Der Klassifikator verwendet die Merkmale zur Klassifikation der einzelnen Artikel, indem er zunächst auf einem Trainingsset trainiert wird und anschließend ein unabhängiges Testset klassifiziert. Das Trainingsset besteht aus 871 und das Testset aus 150 Zeitungsartikeln.

Zunächst soll der Artikel mittel Klassifikation kategorisiert werden. Bei der Klassifikation wird für jeden Artikel die Wahrscheinlichkeit der Zugehörigkeit zu einer Kategorie (*uncertainty*) errechnet. In die Kategorie mit der höheren Wahrscheinlichkeit wird dann der Artikel eingeordnet. Anschließend wird die Genauigkeit (*accuracy)* des BoW-basierten mit der des LDA-basierten Klassifikationsmodells verglichen. Insgesamt besteht das verwendete Korpus aus 1.021 Artikeln und unterteilt sich in 507 (49,7%) relevant und 514 (50,34%) nicht relevant kategorisierte Artikel.

4. Die Testläufe

Für die verschiedenen Testläufe wurde das Zeitungskorpus zunächst von Stoppwörtern bereinigt. Für die Klassifikation wurde ein Maximum Entropie Algorithmus ausgewählt. Dieser Klassifikationsalgorithmus eignete sich sowohl für das BoW- als auch für das LDA-basierte Modell, was verschiedene Testläufe bestätigten.

Da es sich bei LDA um ein auf Wahrscheinlichkeiten basierendes Modell handelt, sind die generierten *topics* nicht völlig identisch reproduzierbar. Um die Vergleichbarkeit zu erhöhen, wurden die *topics* einer manuell festgelegten *topic*-Anzahl in Ausgabedateien gespeichert um in weiteren Testläufen, soweit möglich, verwendet zu werden. Für Testläufe mit verändertem Korpus, zum Beispiel durch Lemmatisierung, mussten neue *topics* generiert werden. Um die Vergleichbarkeit zu erhöhen und das Ergebnis nicht zu stark auf einen exemplarisch generierten Testlauf zu stützen, wurden für jede *topic*-Anzahl jeweils 10 Testläufe durchgeführt, aus welchen ein Durchschnitt errechnet wurde. Alle Ergebnisse sind deshalb Durchschnittswerte aus 10 Testläufen der jeweiligen *topic*-Anzahl. Die Standardabweichung (Standard Deviation) zwischen den Ergebnissen erwies sich allerdings als sehr gering und lag unter 0,0075 (die Ergebnisse sind also sehr stabil). Dies wurde mit einem manuell erstellten Goldstandard als Testset ermittelt.

5. Evaluation der Ergebnisse

Beim detaillierten Vergleich der zehn Testläufe innerhalb einer *topic*-Anzahl zeigte sich der größte Vorteil des LDA-basierten Modells. Basierend auf den manuellen Annotationen wurden die vom System falsch klassifizierten Artikel bestimmt. Die durchschnittliche *uncertainty* dieser Artikel war beim LDA-basierten Klassifikationsmodell näher an 0,5 als bei dem BoW-basierten Modell (vgl. Tabelle 1). Die Wahrscheinlichkeit 0,5 bildet hier die entscheidende Grenze, an der Artikel als relevant oder nicht relevant klassifiziert werden.

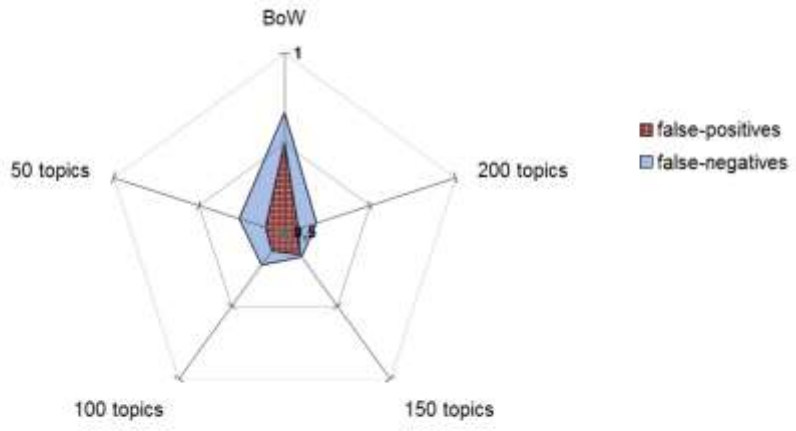


Abbildung 1: Durchschnittliche uncertainty

Bei der Betrachtung der durchschnittlichen *uncertainty*-Werte wird zwischen false-positives und false-negatives unterschieden. Manuell als relevant annotierte Artikel, welche vom System als nicht relevant klassifiziert wurden, sind false-positives. False-negatives beschreibt vom Annotator als nicht relevant bewertete Artikel, welche das System irrtümlich als relevant klassifiziert hat. Abbildung 1 soll dies verdeutlichen.

	false-positives		false-negatives	
	uncertainty	Standard Deviation	uncertainty	Standard Deviation
BoW	0,7529	0,1401	0,8371	0,1516
50 topics	0,5555	0,0390	0,6329	0,1141
100 topics	0,5602	0,0492	0,6089	0,0730
150 topics	0,5735	0,0545	0,5807	0,0665
200 topics	0,5400	0,0346	0,5933	0,0748

Tabelle 1: Uncertainty

Eine nachträgliche Bereinigung des Korpus von Artikeln in einem *uncertainty*-Bereich von 0,5 bis 0,6 würde zwar nicht alle falsch klassifizierten Artikel erkennen, zeigt aber eine erhöhte Standardabweichung. Die BoW-basierte Variante erreicht eine *accuracy* von 0,8178 mit einer Standardabweichung von 0,0342 (vgl. Tabelle 2).dennoch einen Großteil davon abdecken. Ein Erhöhung der *uncertainty* auf beispielsweise 0,7 würde die *accuracy* zwar verbessern, aber gleichzeitig das Korpus stärker einschränken. Anwendungsszenarien, in welchen ein kleines Korpus mit wenigen, irrelevanten Artikeln sinnvoll ist, sind allerdings denkbar.

In weiteren Testläufen wurde das Testset mittels Kreuzvalidierung (Cross Validation) und nicht mit einem Goldstandard generiert. Der Goldstandard lässt trotz sorgsamer Auswahl Kritik zu (vgl. Kappa-Werte, oben), stellte sich allerdings für die detaillierte Betrachtung einzelner Ergebnisse, wie beispielsweise der *uncertainty*-Werte und einzelner Artikel, als geeignet heraus. Die Cross Validation extrahiert zunächst randomisiert Artikel für ein Testset aus dem Gesamtkorpus. Das Trainingsset wurde auf 90 Prozent und das Testset auf 10 Prozent des Gesamtkorpus festgelegt. Die Cross Validation wurde

mit dem Faktor 100 durchgeführt. Die 100-malige randomisierte Auswahl des Testsets soll den Einfluss der einzelnen Testdaten verringern.

	accuracy	***Standard Deviation***
BoW	0,8178	0,0342
50 topics	0,8407	0,0363
100 topics	0,8417	0,0374
150 topics	0,8383	0,0378
200 topics	0,8384	0,0379

Tabelle 2: accuracy

Die Anzahl der *topics* muss für LDA manuell ausgewählt werden. Die verwendeten Zahlen wurden mit Vortests ausgewählt. LDA-basierte Ergebnisse sind deshalb bereits optimiert. Wesentlich kleinere oder größere *topic*-Anzahlen als die hier verwendeten können ein ganz anderes Bild zeigen. Testläufe mit den Größen 50, 100, 150 und 200 zeigten, dass die *topic*-Anzahl das Ergebnis beeinflusst. Während Testläufe mit 50 *topics* nach der Klassifikation eine *accuracy* von 0,8407 im Korpus erreichten, erzielten 150 *topics* nur eine *accuracy* von 0,8383. Eine Proportionalität von *topic*-Anzahl zur erzielten *accuracy* konnte nicht festgestellt werden. Das LDA-basierte Modell erzielte mit 100 *topics* eine *accuracy* von 0,8417 und mit 200 *topics* von 0,8384. Alle vier der ausgewählten *topic*-Anzahlen brachten aber eine Verbesserung der *accuracy* gegenüber dem BoW-basierten Modell (vgl. Abbildung 2).

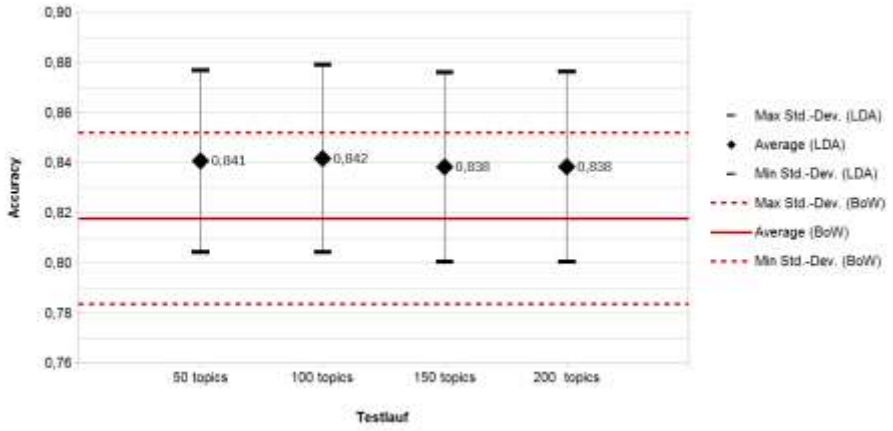


Abbildung 2: Ermittelte accuracy

6. Fazit

Die *accuracy*-Werte des LDA- und BoW-basierten Modells zeigen im Vergleich mit dem Ergebnis der manuellen Annotation, dass dieses für eine textbasierte Klassifikationsaufgabe geeignet sind. Allerdings ist zu beachten, dass LDA durch die manuelle Auswahl der *topic*-Anzahlen bereits optimiert ist. Die Auswahl der *topic*-Anzahl beeinflusst das Klassifikationsergebnis.

Der größte Vorteil des LDA-basierten Modells zeigt sich in der Betrachtung der *uncertainty*. Ein nachträgliches Entfernen der Artikel, welche nur *uncertainty*-Werte im Bereich von 0,5-0,6 erreichen, kann hier zu einer Verbesserung der *accuracy* führen (vgl. Abbildung 1).

Das BoW-basierte Modell zeigte im Vergleich mit den manuellen Annotationen eine Verschlechterung um 0,0222 auf 0,8178. Die *accuracy*-Werte des LDA-basierten Modells bewegten sich für die vier ausgewählten *topics* im Bereich der manuellen Annotationen (zwischen 0,8383 und 0,8417).

7. Literaturverzeichnis

Blei, D. M.; Ng, A. Y.; Jordan, M. I. (2003). Latent Dirichlet Allocation. In: The Journal of Machine Learning Research. Vol. 3, S. 993-1022.

Landis, J. R.; Koch, G. G. (1977). The measurement of observer agreement for categorical data. In: International Biometric Society. Vol. 33(1). S. 159-174. DOI: 10.2307/2529310.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

Universität Stuttgart (2014). Multiple kollektive Identitäten in internationalen Debatten um Krieg und Frieden seit dem Ende des Kalten Krieges. Sprachtechnologische Werkzeuge und Methoden für die Analyse mehrsprachiger Textmengen in den Sozialwissenschaften (eIdentity). Stand: 08.10.2014. Retrieved March 16, 2015 from http://www.uni-stuttgart.de/soz/ib/forschung/Forschungsprojekte/eIdentity.html.

Zhang, R. J. ; Zhou, Z.-H. (2010). Understanding bag-of words model: a statistical framework. In: International Journal of Machine Learning and Cybernetics Vol. 1, S. 43-52.

MusicXML Analyzer

Ein Analysewerkzeug für die computergestützte Identifikation von Melodie-Patterns

Manuel Burghardt, Lukas Lamm, David Lechler, Matthias Schneider und Tobias Semmelmann

Universität Regensburg
Institut für Information und Medien, Sprache und Kultur
Ansprechpartner: manuel.burghardt@ur.de

Zusammenfassung

Der Beitrag beschreibt Aktivitäten aus einem aktuellen Digital Humanities-Projekt, das im Schnittfeld von Informations-, Musik- und Kulturwissenschaft angesiedelt ist. Dabei soll eine Sammlung von ca. 50.000 handschriftlichen Liedblättern mit deutschsprachiger Volksmusik digitalisiert und maschinenlesbar in *MusicXML* kodiert werden, um schließlich über ein Informationssystem quantitative Analysen des Materials zu erlauben. Wir stellen einen ersten webbasierten Prototypen vor, der es erlaubt, Musikstücke im MusicXML-Format statistisch auszuwerten und nach konkreten Melodie-Patterns zu suchen. Das Tool ist zudem in der Lage, virtuelle Partituren und Audioausgaben auf Basis des MusicXML-Markups zu erstellen und direkt im Webbrowser verfügbar zu machen.

Abstract

This article describes a recent digital humanities project by a group of researchers from various disciplines, including musicology, cultural studies, and information science. The project is aiming to digitize a collection of approx. 50,000 handwritten sheets of German folk music. In addition, the songs will be encoded in *MusicXML* format, which makes the data accessible for quantitative analyses. In this article we describe a web-based prototype that can be used to analyze any MusicXML encoded song with regard to quantitative properties (e.g. most frequent notes and intervals), but that also allows researchers to query a corpus of songs for specific melodic patterns. The tool is also designed to render virtual scores and create audio output from the MusicXML markup directly in a user's web browser.

1. Einleitung

Konzepte wie *distant reading* (Moretti, 2007) und *macroanalysis* (Jockers, 2013) beschreiben Ansätze in der Literaturwissenschaft, die es erlauben, auf quantitativer Ebene neue Erkenntnisse zu größeren Zusammenhängen zwischen einzelnen Texten und Autoren zu erlangen. Dabei sollen generische Muster identifizierbar gemacht werden, welche durch *close reading* einzelner Texte nicht unmittelbar sichtbar sind. Es scheint somit naheliegend, quantitative Verfahren, welche größere Datensammlungen zunächst aus der *Distanz* analysieren, auch auf andere Medientypen, wie etwa Audiodaten zu übertragen (für eine beispielhafte Studie vgl. etwa Viglianti, 2007), um so durch *distant hearing* besonders markante oder häufig wiederkehrende Melodie-Patterns in großen Liedkorpora zu entdecken.

1.1 Music Information Retrieval

Das Indexieren und Durchsuchen von Audiodaten fällt gemeinhin in das Feld des *Music Information Retrieval* (MIR). Bestehende MIR-Tools lassen sich einerseits in Tools für die Verarbeitung von Audiodaten (akustische Aufnahmen) und andererseits in Tools für die Verarbeitung von notierter Musik (Transkriptionen / Notenblätter) klassifizieren (vgl. Typke, Wiering, & Veltkamp, 2005). Über einen Vergleich von 17 bestehenden Tools identifizieren Typke et al., (2005) die folgenden zentralen Komponenten für MIR-Systeme, die jeweils auf unterschiedliche Arten umgesetzt werden können:

- *Input*: Formulierung einer Query, entweder im Audioformat (vgl. etwa „Query by Humming“, Ghias et al., 1995) oder als formalisierte, maschinenlesbare Suchanfrage
- *Matching*: Abgleich der Query mit der Musikdatenbank (exakt vs. angenähert, monophon vs. polyphon)
- *Features*: Parameter, welche bei der Formulierung der Query gesetzt werden können (Tondauer, Tonhöhe, Intervalle, u.a.)

Konkrete Ansätze zur quantitativen Analyse von Musikdaten mithilfe von MIR-Systemen finden sich zahlreich in der Literatur, und werden dort meist unter dem Gesichtspunkt der *melodic similarity* beschrieben (Grachten, Arcos, & de Mántaras, 2002; Grachten, Arcos, & Mántaras, 2004; Miura & Shioya, 2003; Müllensiefen & Frieler, 2004a, 2004b).

1.2 Ziele

In diesem Beitrag präsentieren wir ein prototypisches Tool das es erlaubt, beliebige Musikdaten, die im standardisierten *MusicXML*-Format[4] vorliegen, zu analysieren und nach Melodie-Patterns zu durchsuchen. Durch seine Flexibilität hinsichtlich der analysierbaren Daten füllt das Tool damit zum einen eine Lücke in der Landschaft bestehender MIR-Tools, die häufig für eine abgeschlossene Sammlung von Musik konzipiert sind (vgl. Typke et al., 2005). Zum anderen liefert das Tool einen ersten Prototyp für die Analyse von mehreren tausend handschriftlichen Liedblättern, die in einem aktuell laufenden Projekt digitalisiert und im MusicXML-Format kodiert werden. Gleichzeitig illustriert das Tool die Vorteile einer standardisierten, XML-basierten Kodierung von Musikdaten, welche es nicht nur erlaubt die Daten automatisch auswertbar und durchsuchbar zu machen, sondern im Umfeld webbasierter Technologien zudem eine grafische Ausgabe in Form virtueller Partituren sowie eine akustische Wiedergabe der Daten in einer integrierten Analyseumgebung ermöglicht.

2. Projektkontext: Regensburger Volksmusik-Portal

Das in diesem Beitrag präsentierte Analysewerkzeug ist im Kontext eines laufenden Digital Humanities-Projekts, im Schnittfeld von Informationswissenschaft, Musikwissenschaft und Kulturwissenschaften, angesiedelt. Wesentliche Ziele dieses interdisziplinären Projekts sind die Digitalisierung, sowie auch die computergestützte Analyse eines Korpus mit über 50.000 Liedblättern, die zum Bestand des Regensburger Volksmusik-Portals (RVP)[5] zählen. Die einzelnen Liedblätter sind jeweils handschriftlich verfasste Dokumente, welche sowohl die monophonen Melodien als auch die Liedtexte zu Volksliedern aus dem deutschsprachigen Raum im 19. – 20. Jhd. enthalten (vgl. Abbildung 2). In einer vorhergehenden Projektphase wurden die Metadaten (Sangesort, Incipet, etc.) des Bestands bereits größtenteils in ein digitales Archivsystem überführt.

---

[4] Mehr Informationen zum MusicXML-Standard finden sich auf der offiziellen Webseite http://www.musicxml.com/. *Hinweis*: Sämtliche URLs die in diesem Beitrag erwähnt werden wurden zuletzt am 12.6.2015 auf Verfügbarkeit geprüft.

[5] http://www.uni-regensburg.de/bibliothek/projekte/rvp/index.html

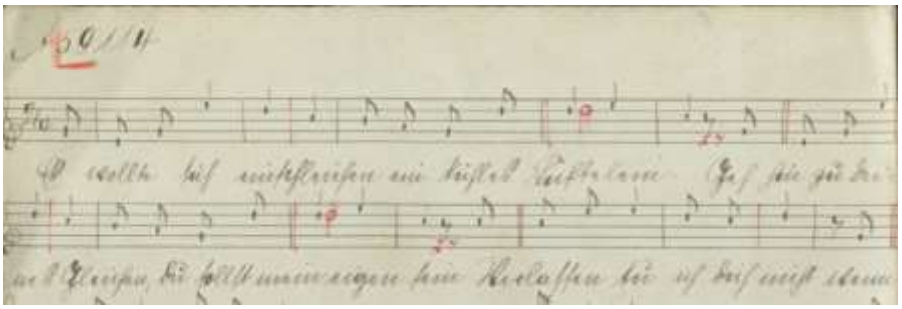


Abbildung 2: Teilausschnitt eines Liedblattes aus der RVP-Sammlung.

Im Rahmen des Projekts soll u.a. diese bestehende Liedblattsammlung digitalisiert und über ein Webportal öffentlich verfügbar gemacht werden. Dabei sollen die Daten nicht nur gescannt werden, sondern so repräsentiert werden, dass quantitative Analysen des Materials möglich werden, also z.B. die Identifikation typischer Melodie-Fragmente in bestimmten geografischen Bereichen oder zeitlichen Epochen. Aktuell wird geprüft, inwieweit mit automatischen *Optical Music Recognition* (OMR)-Verfahren (Bainbridge & Bell, 2001; Raphael & Wang, 2011; Rebelo, Capela, & Cardoso, 2010) die handschriftlichen Notenblätter in das etablierte MusicXML-Format überführt werden können. Da die ersten Evaluationsergebnisse zu solch automatischen Verfahren bislang wenig aussichtsreich für die vorliegende Datenbasis scheinen, wird alternativ angedacht, eine webbasierte Crowdsourcing-Plattform nach dem Vorbild bestehender Transkriptionsprojekte, wie etwa „What's the Score"[6], umzusetzen.

3. MusicXML Analyzer

MusicXML Analyzer unterstützt Kultur- und Musikwissenschaftler dabei, Musik einerseits statistisch auszuwerten und andererseits wiederkehrende Melodie-Patterns in unterschiedlichen Musikstücken aufzufinden. Vor dem Hintergrund der besonderen Anforderungen an die *Humanist-Computer Interaction* (Burghardt & Wolff, 2015) wird beim Interface-Design, d.h. beim Upload von Dokumenten, der Eingabe von Suchmustern sowie der Darstellung der Auswertung und der Ergebnisse, besonderes Augenmerk auf *Usability* und *User Experience* gelegt. Die Ergebnisse der Analysen werden direkt im Browser visualisiert und können darüber hinaus auch als Audioausgabe abgespielt werden. Daneben ist aber auch der Export der Analysen als PDF

[6] http://www.bodleian.ox.ac.uk/bodley/finding-resources/special/projects/whats-the-score

oder als CSV-Datei (*Comma Separated Values)* zur weiteren Verarbeitung der Daten möglich.

3.1 Verfügbarkeit

Eine erste, funktionsfähige Version des *MusicXML Analyzer* ist unter folgender Adresse verfügbar, und kann direkt im Browser ausprobiert werden[7]:

- Demo: http://music-xml-analyzer.herokuapp.com/

Gleichzeitig ist das Tool mit allen Komponenten als *Open Source Software* auf dem Code Repositorium *GitHub* verfügbar und kann bei Bedarf beliebig angepasst und weiterentwickelt werden:

- GitHub Repository:
  https://github.com/freakimkaefig/Music-XML-Analyzer

Um einen ersten Eindruck über die wesentlichen Funktionen des Tools zu gewinnen, wurde zudem ein kurzes Demonstrationsvideo erstellt:

- Video: https://www.wevideo.com/view/417278189

3.2 Funktionsübersicht

Die Funktionen des Systems lassen sich in drei Teilbereiche aufgliedern, die im Wesentlichen den typischen Analyse-Workflow widerspiegeln (vgl. Abbildung 3). Im ersten Schritt erfolgt der Upload von Dateien im MusicXML-Format, daraufhin folgt die automatische, statistische Auswertung. Den letzten Schritt stellt die Suche nach Mustern in den hochgeladenen Dateien dar.

[7] Hinweis: Für die aktuell verwendete Serverumgebung gelten einige technische Einschränkungen: Bei längerer Inaktivität der Anwendung begibt sich die Laufzeitumgebung automatisch in einen Ruhemodus, welcher das erneute „Aufwecken“ der Anwendung erfordert. Das führt dazu, dass der Initialzugriff auf die Anwendung unter Umständen einige Sekunden dauern kann.

Abbildung 3: Startscreen der Anwendung, auf welchem der grundlegende Tool-Workflow (Datei-Upload > Analyse > Mustersuche) dargestellt wird.

3.2.1 Upload

Bevor die Analyse von Musikstücken beginnen kann, müssen zunächst Dateien im MusicXML-Format hochgeladen werden. Der Upload beliebig vieler MusicXML-Dateien kann über einen intuitiven Drag-and-Drop-Dialog bewerkstelligt werden. Die Anwendung gibt jeweils Rückmeldung über den aktuellen Upload-Status, da der Prozess bei mehreren Dateien und einer langsamen Internetverbindung einige Sekunden in Anspruch nehmen kann.

3.2.2 Analysekomponente

Nach erfolgreichem Upload erfolgt die Analyse der MusicXML-Dateien. Zum Parsing der MusicXML-Daten kommt eine Kombination aus XPath-Ausdrücken und PHP-Skripten zum Einsatz, die serverseitig ausgeführt werden und die Ergebnisse der Analyse in einer relationalen SQL-Datenbank speichern. Diese Parsing-Ergebnisse stellen einerseits die Grundlagen für die spätere Mustersuche dar, und werden andererseits (in Teilen) in einem Dashboard visualisiert (vgl. Abbildung 4).

Abbildung 4: Ausschnitt aus dem Analyse-Dashboard
für ein Korpus aus MusicXML-kodierten Dokumenten.

Die folgenden Informationen sind über das Dashboard sowohl für das gesamte Korpus als auch für jeweils einzelne Lieder des Korpus verfügbar:

- Anzahl aller Einzelnoten, Pausen und Takte
- Verwendete Instrumente (sofern in MusicXML angegeben)
- Häufigkeitsverteilung von Notenwerten, Intervallen, Tonarten, Notenlängen und Takten

Alle im Dashboard dargestellten Analysen können zudem als CSV-Datei heruntergeladen werden.

3.2.3 Such-Komponente

Die Such-Komponente erlaubt es schließlich das Korpus aus MusicXML-Dokumenten nach bestimmten Melodie-Patterns zu durchsuchen. Dabei ist die Suche nach reinen Rhythmus-Patterns, reinen Tonabfolgen sowie einer Kombination aus Rhythmus- und Ton-Patterns möglich (vgl. Tabelle 1):

Parameter	Beispielhafte Suchanfrage
Tonabfolge-Pattern (*sound sequence*)	Finde alle Songs in denen die Notenfolge C - C - D - G (jeweils 4. Oktave) vorkommt, ungeachtet der Notenlänge.
Rhythmus-Pattern (*rhythm*)	Finde alle Songs in denen eine Sequenz aus folgenden Notenlängen vorkommt, ungeachtet der Tonhöhe: Achtelnote - Achtelnote - Sechzehntelnote - Sechzehntelnote.
Melodie-Pattern (*melody*), d.h. Kombination aus Tonwert und Rhythmus	Finde alle Songs in denen die Notenfolge C, Achtelnote - C, Achtelnote - D, Sechzehntelnote – G, Sechzehntelnote (jeweils 4. Oktave) vorkommt.

Tabelle 1: Suchparameter und beispielhafte Suchanfragen.

Die Eingabe der Suchmuster ist entweder durch Klicken mit der Maus über eine interaktive Notenzeile möglich, oder über die Auswahl von Noten- und Taktwerten über ein Menü (vgl. Abbildung 5). Die ausgewählten Melodie-Patterns werden jeweils direkt auf der Notenzeile dargestellt und können zudem im Browser abgespielt werden, um auch einen direkten, akustischen Eindruck der formulierten Suchanfrage zu erhalten.

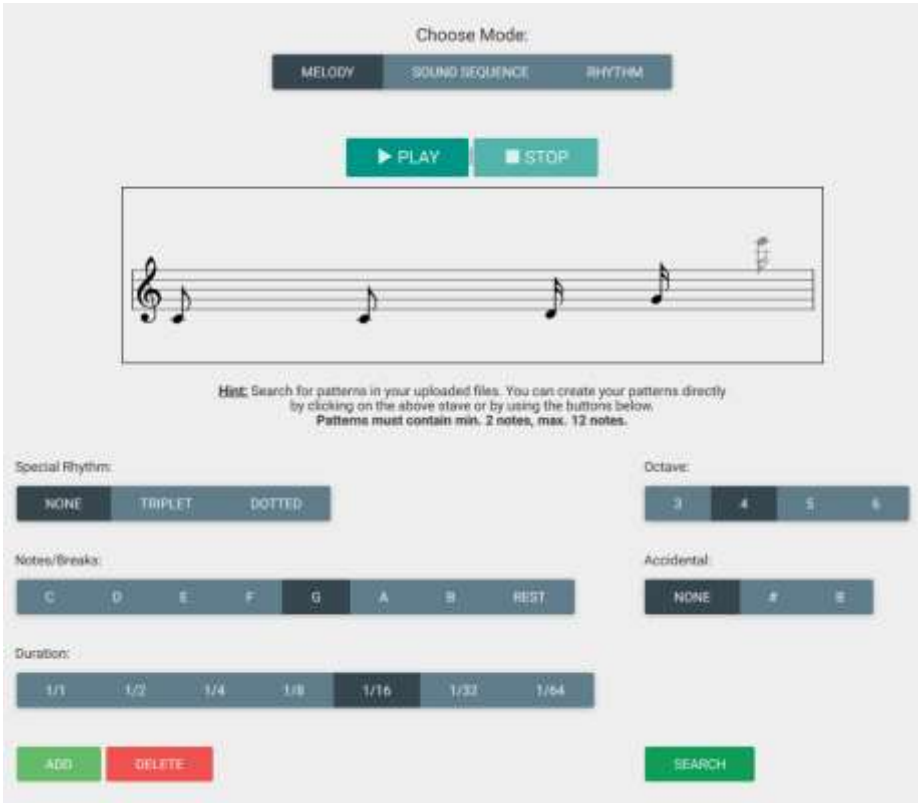


Abbildung 5: Interface für die Formulierung von Suchanfragen
zur Identifikation von Tonfolge-, Rhythmus- oder Melodie-Patterns.

Nach dem Abschicken der Suchanfrage werden in einer Ergebnisliste all diejenigen Lieder angezeigt, in denen das Such-Pattern vorkommt, zusammen mit der jeweiligen Häufigkeit des Vorkommens. Mit einem Klick auf eines der aufgelisteten Lieder wird automatisch die entsprechende Partitur anhand der MusicXML-Daten im Browser rekonstruiert und die gesuchte Sequenz darin farbig hervorgehoben. Auch hier ist es wiederum möglich, die gesamte Partitur im Browser abzuspielen oder diese als PDF herunterzuladen (vgl. Abbildung 6).

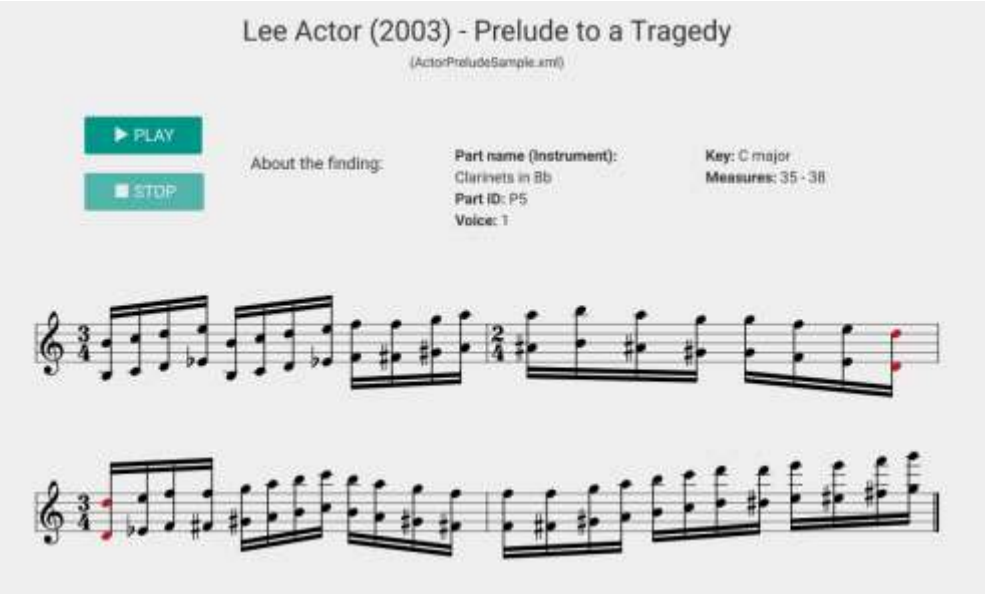


Abbildung 6: Virtuelle Partitur-Darstellung des Trefferdokuments
und farbige Hervorhebung der gesuchten Notensequenz.

4. Implementierung

4.1 Programmbibliotheken und Webtechnologien

MusicXML Analyzer wurde ausschließlich mithilfe von Webtechnologien wie HTML, CSS, JavaScript (JS) und PHP umgesetzt. Für die Implementierung der einzelnen Komponenten wurde, soweit möglich, auf bestehende Frameworks und Programmbibliotheken zurückgegriffen (vgl. Tabelle 2).

Funktion	Name	URL
Allgemeines PHP Framework	Laravel	http://laravel.com/
Allgemeines JavaScript Framework	jQuery	https://jquery.com/
Allgemeines CSS Framework	Bootstrap	http://getbootstrap.com/
JS-Bibliothek für die Visu-	D3.js	http://d3js.org/

alisierungen im Dashboard		
JS-Bibliothek für die Erzeugung der virtuellen Partituren	Vexflow	http://www.vexflow.com/
JS-Bibliothek für die Erzeugung der Audioausgabe	Midi.js	http://mudcu.be/midi-js/
JS-Bibliothek für die Visualisierung von Statusnachrichten	Typed.js	http://www.mattboldt.com/demos/typed-js/
JS-Bibliothek für den Upload von Dateien	Dropzone.js	http://www.dropzonejs.com/
JS-Bibliothek für den Export von PDFs	jsPDF	https://parall.ax/products/jspdf

Tabelle 2: Überblick zu den verwendeten Frameworks und Programmbibliotheken.

4.1.1 Systemarchitektur und Implementierungsdetails

Abbildung 7 zeigt den Aufbau des verwendeten Laravel-Frameworks im Zusammenspiel mit den jeweiligen JavaScript-Komponenten. Die Laravel-Komponente kümmert sich dabei um die serverseitige Logik der Anwendung, wie etwa die Auslieferung von HTML-Seiten über eine integrierte *templating engine* oder die Kommunikation mit der Datenbank zur persistenten Speicherung. In Abbildung 7 wird auch der Aufbau nach dem Model-View-Controller-Pattern (MVC) deutlich, demzufolge die grafische Repräsentation (*view*) von der Logik (*controller*) und der Datenschicht (*model*) getrennt ist. Die JavaScript-Komponente der Anwendung ist ebenso nach dem MVC-Pattern aufgebaut, und kümmert sich hauptsächlich um die Interaktivität der Anwendung. Beispielsweise werden die Suchmuster in JavaScript zunächst clientseitig zwischengespeichert und erst beim Absenden der Suchanfrage an den Server geleitet. Weitere Aufgabengebiete sind die Visualisierung der Analysedaten und Partitur-Ausschnitte, sowie die Audio-Wiedergabe einzelner Sequenzen.

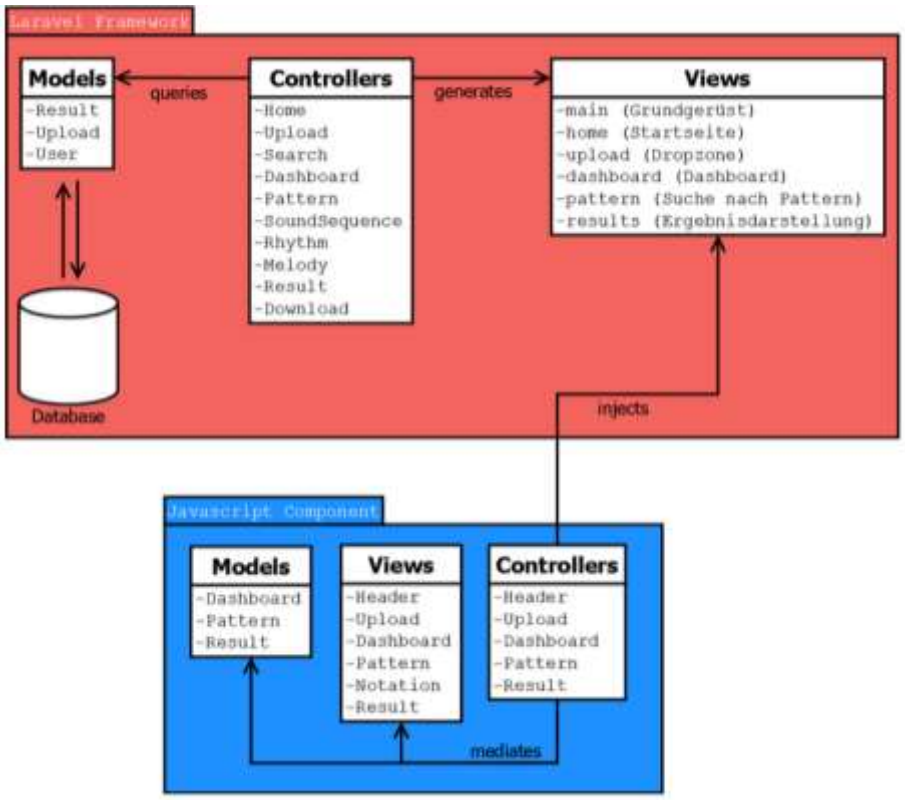


Abbildung 7: Überblick zur Systemarchitektur auf Basis des PHP Laravel-Frameworks.

4.1.2 Datenspeicherung und Benutzerverwaltung

Bei der Konzeption wurde bewusst auf eine aufwendige Nutzerverwaltung mit Registrierung und Anmeldung verzichtet. Um dem Nutzer die Möglichkeit zu bieten, seine Arbeit ohne zusätzliche Speicherung immer aktuell zu halten werden alle Analysen automatisch in einer Datenbank gespeichert. Den Nutzern wird bei der ersten Verwendung des Tools automatisch eine ID zugewiesen, die gleichzeitig als Cookie im Browser hinterlegt wird. Auf den Server geladene Dateien werden entsprechend mit dem Nutzer referenziert. Nach dem gleichen Schema werden einem Upload-Objekt nach erfolgreicher Analyse dessen Ergebnisse zugeordnet.

5. Ausblick

MusicXML Analyzer hat aktuell den Status eines voll funktionsfähigen Prototypen, der in dieser Form direkt in der verfügbaren Liveversion oder als lokale Webapplikation für eigene Musikanalysen verwendet werden kann.

Gleichzeitig wird die Anwendung während des eingangs beschriebenen Projekts stetig weiterentwickelt. Parallel dazu sind Weiterentwicklungen und Anpassungen des Tools durch andere Entwickler jederzeit möglich, da das Tool auf *GitHub* unter der freien *MIT Lizenz* als *Open Source Software* veröffentlich wurde.

Ein unmittelbar nächster Schritt für die Erweiterung von *MusicXML Analyzer* im Rahmen des laufenden Volksmusikprojekts ist die Verknüpfung der Datenbank mit den bereits digital vorliegenden Metadaten (Ort, Datum, u.a.), sodass etwa eine gezielte Suche nach Melodie-Patterns in bestimmten Regionen oder Zeiträumen möglich ist. Während im vorliegenden Prototyp zunächst die explizite Suche nach Patterns im Vordergrund steht, so soll parallel auch eine Analysekomponente erstellt werden, die in den MusicXML-kodierten Daten automatisch melodische Ähnlichkeiten und Muster erkennt (vgl. hierzu auch die eingangs zitierten Studien im Bereich *melodic similarity*). Neben der reinen Analyse von Melodie-Patterns ist auf lange Sicht zudem die Kombination mit den digitalisierten Liedtexten ein wichtiges Desideratum, welches es beispielsweise erlaubt zu untersuchen ob bestimmte Wörter mit bestimmen melodischen Charakteristika korrelieren. Eine beispielhafte Fragestellung könnte etwa folgendermaßen lauten: Kommt das Wort „Krieg“ wegen der negativen Konnotation häufiger in Kombination mit einer Molltonart als mit einer Durtonart vor?

6. Literaturverzeichnis

Bainbridge, D., & Bell, T. (2001). The challenge of optical music recognition. *Computers and the Humanities* (*35)*, 95–121.

Burghardt, M., & Wolff, C. (2015). Humanist-Computer Interaction: Herausforderungen für die Digital Humanities aus Perspektive der Medieninformatik. Book of Abstracts Workshop “Informatik und die Digital Humanities”, Leipzig.

Grachten, M., Arcos, J. L., & de Mántaras, R. L. (2002). A comparison of different approaches to melodic similarity. Proceedings of the 2nd ICMAI, p. 1-12.

Grachten, M., Arcos, J. L., & Mántaras, R. L. De. (2004). Melodic Similarity: Looking for a Good Abstraction Level. Proceedings of the 5th ISMIR, p. 210-215.

Ghias, A., Logan, J., Chamberlin, D., & Smith, B.C. (1995). Query By Humming – Musical Information Retrieval in an Audio Database. Electronic Proceedings of the ACM Multimedia 95.

Jockers, M. L. (2013). Macroanalysis: Digital Methods and Literary History (Topics in the Digital Humanities). University of Illinois Press.

Miura, T., & Shioya, I. (2003). Similarity among melodies for music information retrieval. Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM '03), p. 61-68.

Moretti, F. (2007). Graphs, Maps, Trees: Abstract Models for Literary History. London: Verso.

Müllensiefen, D., & Frieler, K. (2004a). Melodic Similarity: Approaches and Applications. Proceedings of the 8th International Conference on Music Perception & Cognition, p. 283–289.

Müllensiefen, D., & Frieler, K. (2004b). Optimizing Measures Of Melodic Similarity For The Exploration Of A Large Folk Song Database. Proceedings of the 5th ISMIR, p. 274–280.

Raphael, C., & Wang, J. (2011). New Approaches to Optical Music Recognition. Proceedings of the 12th ISMIR, p. 305–310.

Rebelo, a., Capela, G., & Cardoso, J. S. (2010). Optical recognition of music symbols. *International Journal on Document Analysis and Recognition* (*13)*, 19–31.

Typke, R., Wiering, F., & Veltkamp, R. C. (2005). A survey of music information retrieval systems. *Proceedings of the 6th ISMIR*, p. 153–160.

Viglianti, R. (2007). MusicXML: An XML Based Approach to Musicological Analysis. *Proceedings of the 18th Digital Humanities conference*, p. 235–237.

Session 2
Information Retrieval, Textmining und Informationsverhalten

Outliers are the better participants

Elke Greifeneder

Institut für Bibliotheks- und Informationswissenschaft
Humboldt-Universität zu Berlin
Unter den Linden 6, 10099 Berlin
greifeneder@ibi.hu-berlin.de

Abstract

Outliers are participants whose scores are too extreme to be included in statistical analyses and are, hence, considered to be lost participants. In information behavior publications, descriptions of outliers and their handling are sparse or non-existent. This contribution suggests that outliers are not necessarily lost participants, but may, on the contrary, be an indication of normal events in the natural environment and thus an indispensable key to interpret information behavior. In this sense, outliers could even be better representations of real users than the statistically standard participant. Data for the outlier analysis come from an asynchronous remote user test that examined simple known-item search task completion in a laboratory and a natural environment test setting.

1. Introduction

Information behavior examines how people interact with information (Bates, 2010). Users' information context has been raised as an important factor for many years (Johnson, 2003; Kelly, 2006; Ingwersen, 2007), but only since the increase of mobile information behavior and the expected uptake of smart environments, i.e. environments that respond and anticipate a need, context has gained a new importance in information behavior.

Contextual influences can be visible in outlying scores. Outliers are participants whose scores are too extreme to be included in statistical analyses and are, hence, considered to be lost participants. In information behavior publications, descriptions of outliers and their handling are sparse or non-existent. The usual way of dealing with outliers is to exclude them from the dataset. Outliers indicate discrepancies in a data set and, in general, researchers wish to dispose of them as soon as possible. Approaches to outliers vary. Some researchers delete outliers before running their analysis; others replace them with mean or median values. In both cases researchers often declare that their data follow a normal distribution and that the problematic values were exceptions to their model. A more rigorous statistical approach may, however, be

to consider whether the outliers do not properly reflect the true situation and to examine what exactly they mean in terms of the model.

This article discusses whether the outliers in an asynchronous remote user test of five digital libraries (Greifeneder, 2015) were improper values or whether they reflect behavior of regular users. Data from a search-task evaluation study in an experimental setting in both a laboratory and in participant's natural environment serve as the source.

2. Background

Outliers are rarely studied in depth in information science. In fact, a literature analysis revealed only two studies in this field: Kelly & Peacock (1999) examined trends in web service provision and discovered that the separate analysis of outliers provided information, which was not found by the initial robot technic. Miller et al (2014) examined outliers in order to detect spam in Twitter tweets.

Grubbs' work (1996), which is still the reference for research on outliers, defines two variations of outliers: an "outlying observation may be the result of gross deviation from prescribed experimental procedure or an error in calculating or recording the numerical value" (p. 1). Grubbs describes the second variation as follows: "An outlying observation may be merely an extreme manifestation of the random variability inherent in the data. If this is true, the values should be retained and processed in the same manner as the other observations in the sample" (Grubbs, 1996, p. 1). This means that an outlier might be only an extreme manifestation within the data set and not a value outside of it. Deleting this kind of outlier means concealing and thus distorting the real distribution. Grubbs' quote contains another important point about the variability inherent in the data. If the detected outliers are not errors, then data's variability may be higher than expected. That is, if the outliers are not just errors, the outlier's score must be treated as genuine.

It is unusual that a sample follows a normal distribution without outlying scores and, thus, it is expected that all information behavior studies using inferential statistics mention outliers. This turned out to be a wrong assumption. An analysis of publications in the Information Science category in *Science Direct* and *Emerald Insight* using the full text query term "outlier" produced fewer than 10 studies who report on outliers. Only a small number of studies in information science treat outliers as genuine and publish information on outliers and how these have been treated. The most detailed analysis is offered by Goodall (2006) who tested statistically the influence of low outlying numbers on the overall result. Huuskonen & Vakkari (2008) explain

that they had to discard four of 46 questionnaires (9%) "due to their incorrect [Medical subject headings] term explosions, which produced hundreds of terms" (p. 291). Breitbach & Prieto (2012) similarly study more closely the reason behind the extreme values and discover that the outlying participants were patrons outside of library opening hours. Tenopir et al (2013) mention "over 150 outliers" from a dataset of 1,078 responses (13%). Their reasoning of excluding outliers was to "achieve a more representative average, rather than allowing a significantly higher number to skew the data" (p. 979). Lower numbers of outliers are presented by MacFarlane et al (2010) who offer information on only one user, a dyslexic user who "used the internet for 50 hours p/w" (p. 982) and by Kelly et al (2008) who explain that they had originally 52 subjects and needed to exclude one, because this subject was considered an outlier, since the "subject's age was over seven standard deviations above the mean" (p.128). Sewell (2013) mentions one outlier who tweeted 36,922 times and was obviously an extreme – but genuine – Twitter user.

Nielsen (2006) analyzes a large quantitative data set for differences in behavior between males and females and discovers that of "1,520 cases, eighty-seven were outliers with exceedingly slow task times. This means that 6% of users are slow outliers". Instead of deleting the outliers, Nielsen examines them more closely and concludes that "slow outliers are caused by bad luck rather than by a persistent property of the users in question" (online source). What he calls "bad luck" is a test situation, in which a participant might overlook a link, or got on the wrong track and was unable to find the answer. Nielsen also makes clear that these outliers are unwelcome for statistical tests, but are too many to be thrown out of the analysis.

Another group of publications report outliers, but refrains from describing them. Johansen et al. (2011) report – without further details – on the outliers of a usability study that "401 samples out of the 57,600 were considered outliers and removed from the analysis" (p. 1181). Similarly Gao (2004) explains that "[b]oth models had good fit after removing outliers in data points" (p. 973) and Heidar et al (2013) state that "[w]hen the outliers in the data were eliminated, the [need for cognition] effect became statistically significant" (p. 969). Joho et al (2015) state that the difference "between Past/Future and Recency appears to be relatively large with some outliers" (p. 11) without explaining what caused these data points to be outliers and how they were treated. Since the authors state in the following sentence that an ANOVA revealed a particular result, it is to be assumed that the outliers were discarded. Finally, Wijaya & Bressan (2006) study clustering of web documents and reach the conclusion that "17 out of 90 root set documents labeled in categories pertaining to the query 'pyramid' are not grouped and hence are outliers" (p. 986). In Wijaya & Bressan's opinion, the extreme

values are part of what Grubb called the first variation and can therefore be discarded. However, if the extreme values belong instead in the second variation group and are an extreme manifestation of the random variability inherent in the data, the conclusion could be different. It might be possible that the 17 documents are not outliers, but manifestations that the model does not fit. Discarding outliers as not fitting values is a danger to theory development in information science. The lack of reporting outliers at all is an even bigger danger to the credibility of information science research.

3. Data description

In the following part, the method for data collection, asynchronous remote usability tests, is explained, followed by the study setup, recruiting and participants demographics.

Asynchronous remote usability tests allow users to participate in a study without being restricted by territorial and/or temporal constraints. The tests can be accessed anywhere and at any time (Albert et al., 2010) and participants can access a digital service and conduct tasks, like in a thinking-aloud test. Researchers can follow the participants' path through click tracking and can ask questions. Researchers learn what number of participants solved, and how many abandoned a task. They also learn how much time participants needed to complete a specific task.

In order to study the impact of distraction on user's information behavior in a natural environment setting, an experiment was designed in which one group of participants completed an asynchronous remote usability test in a laboratory and another group completed it in their own natural environment. The recruitment took place in the lobby of the Jacob-und-Wilhelm-Grimm-Zentrum, which is the main university library building of the Humboldt-Universität zu Berlin. The laboratory was a small computer pool inside the university library building and easily accessible. Participants in the laboratory were not allowed to talk or to distract themselves in any obvious way. There was always at least one recruiter watching and standing guard over the participants in the laboratory. The remote situation was completely different: there was no specification and no demand where and when the test should be accomplished. Recruiters told participants that they can do the test whenever and wherever they wanted and on a machine of their choice. Participants received a link on a sheet of paper and were asked to go to the website and participate in the test. They were not given any restrictions: for example they had no instructions to close any running applications and no ban on talking to others. The aim was to have a realistic remote test environment that resembled the user's everyday use environment. The experiment examined the

impact of distraction on completion scores. In total, 41 students participated in the laboratory and 43 in the remote setting (including outliers).

Participants had to complete small search tasks in five digital libraries in the following order: Perseus, Social Science Open Access Repository (SSOAR), Digital Picture Archives of the Federal Archives (Bundesarchiv), Valley of the Shadow and Amazon.de. Perseus is a digital library of art and archaeology images, *SSOAR* is a full text server for social science publications, *Bundesarchiv* offers access to pictures of the collection of the Federal Archive, *Valley of the Shadow* offers material on US political events that occurred between 1859 and April 1870 and *Amazon.de* is an online shop.

Participants had to search for a concrete article, for a specific picture, for a book and for a page number. If participants found the requested information, they could click on "task completed", or if not, on "abandon the task". They could also click on "task completed" without the relevant information. Their answer was then marked in the data as a task failure. The participants were told that the aim of the study was to improve the usability of digital libraries. They did not know that their time was being measured, or that the study was about their distraction level in a natural environmental setting. The questions regarding whether the participants felt distracted came at the end of the test.

During the test, different types of data were collected. For this test, the software *Loop11* was used. It provides information about the time-and-task-performance such as how much time users needed and how many page views were necessary to complete a task. A click stream record of each participant was also provided. In addition, participants were asked demographic questions and, more importantly, questions on the context of their usage: whether they used other applications during the test and if yes, how often; whether someone contacted them and how they rated their overall concentration on the test. Participants were also asked if any digital library had technical problems during the test period. Different time scores were collected during task completion and during the post-task questionnaire. The following scores were collected that are relevant for this analysis:

- *Test duration*: time spent to complete the tasks and the post-task questionnaire
- *Completion time in X*: time to complete the task in digital library X
- *Page views*: average number of clicks needed to complete all tasks
- *Page views in X*: number of clicks needed to complete task in X

A more detailed description of the test setting, the test description, test results as well as the relevant research background can be found in Greifeneder

(2015). The study showed that participants in the natural environment took longer to complete the same test than the users in the laboratory, but that they were as successful in task completion and needed a similar amount of page views in both settings.

4. Outlier analysis[8]

During the data analysis, it became obvious that the data did not meet the standards of normality. Most statistical tests like the t-test or Pearson's product-moment correlation coefficient assume normality on the distribution of scores on a dependent variable and test results must be treated very carefully when not meeting these requirements. A first boxplot (figure 1) shows that there were two extreme points in the natural environmental setting marked with asterisks (Participants 20 and 42) and a few other outliers in both settings. The mean of the variable *test duration* calculated with both extreme outliers is 32 minutes and without the two participants, it is 22 minutes.

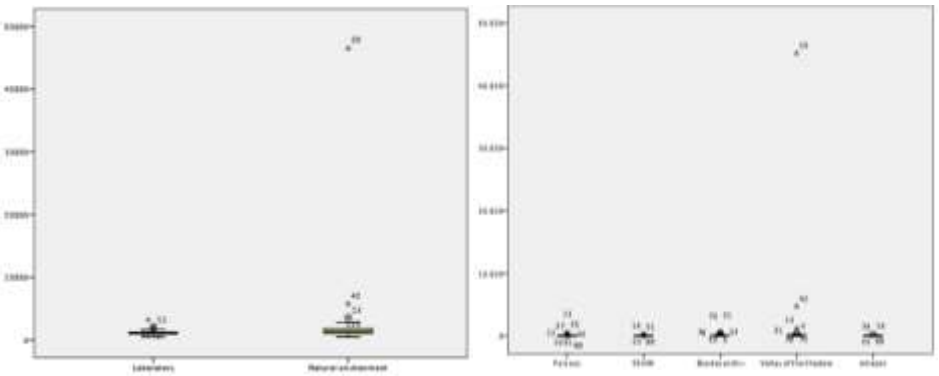


Figure 1 (left). Outliers number 20 and 42 on variable *test duration* (y-axis)
Figure 2 (right). *Completion time in X* (y-axis) including outliers.

It is evident that the two extreme points might be genuine participants, but will have to be discarded before running any statistical tests. Before this step, it is worthwhile to look more closely at the two individual participants and their test behavior. What makes them outliers? Figure 2 shows the boxplots of the variables *completion time in X*, which measures the time spent on the task in each of the digital libraries (natural environment and laboratory). In the following analysis of the two outlying time scores, these scores are compared to the mean of the group, that is M = 22 min.

[8] This analysis has been previously published in parts in Greifeneder (2012) as part of a dissertation project, which is published in Open Access.

The figure illustrates that both participants are only marked as outliers for the task in the digital library *Valley of the Shadow*. This means, they have met the normal distribution for this sample, except for this particular task. Following Nielsen's suggestion, they must have had "bad luck". In order to investigate what else might have happened during the test, the additional information gathered in the test provides a useful source.

Participant 20 was a young female completing the test at home. She needed roughly 12 hours to complete the test. Obviously, this number seems to be unrealistic. However, it is no error in *SPSS* or a data export error. It's possible, of course, that an error on the software side occurred whilst logging that particular participant. The young female needed in average 4.5 page views to complete a task, which is even faster than the mean 5.9 page views of the whole group. She completed all tasks as fast as the average participant and completed all tasks but *Valley of the Shadow* successfully. Only for the task in this specific digital library she needed 45,213 seconds (compared to the mean of 210 seconds). She clicked "task complete" without the right result after four page views (the mean score for *Valley of the Shadow* was 11.2 page views*)*. Additional information reveals that she was using her iPhone to complete the test. She admits that she had several applications open during the test and that she looked at least five times at the applications. She appears to be a multi-tasker. She also admits that apart from being distracted by other applications, she was contacted (either by phone, SMS, or in person) during the test. She admits that she was disturbed by these contacts, but the distraction was not excessive. If the strange number of 12 hours spent on the task in *Valley of the Shadow* is not an error, it also might be that the participant was disturbed during the test and then forgot about the test for several hours and discovered the open test window later and decided to go on. This might be an unusual test behavior, but not an unrealistic one.

Participant 42 was also a female who did the test at home. She needed about an hour and a half for the whole test. The mean score to complete the test was 22 minutes. As in the first case, this participant did not need many page views to complete tasks and was in general successful (her average page views per task was 6.8 and the mean score of the whole group was 5.9 page views). Again, she needed most of the time on the task in *Valley of the Shadow* (80% of the time for the whole test), but needed also many more page views (24 page views compared to the mean of 5 page views). The data state that she abandoned the task. In her comments she wrote that she tried to complete the task, but could not find the requested document. She obviously falls into Nielsen's category of "bad luck". She had one other application open during the test, but claims not to have been distracted by it. She admits a very strong disturbance by someone who contacted her during the test. It is hard to tell without further contextual data whether mere "bad luck" made

her be so slow or whether the disturbance had a significant effect on her time score. Participant 42 volunteered additional information: when asked if any technical problem occurred during the test, she stated that the digital library *SSOAR* did not load. This was surprising news, since *SSOAR* is a trusted repository that should be accessible at all the times and should not have loading problems. In the case of participant 42, this additional information about server problems can help in understanding her behavior in this particular digital library: she needed 136 seconds (M = 73 seconds), but only one page views (M = 3.6 page views). She clicked on "task complete" with all other tasks, even with the more complicated task in the *Bundesarchiv*, but abandoned *SSOAR* which was one of the easiest tasks. Obviously, she was inculpable, because due to the server outage she was unable to do the task. It would be hard to explain strange numbers in online tests without intentionally gathering a minimum of context information, and information about technical problems appears to be a particularly valuable variable to collect. Deciding what to do with participant 42 is difficult. She could not complete a task and needed much time, because she had several forms of "bad luck": she was disturbed and one of the digital libraries did not load, so that she was unable to complete the task. It is tempting to treat her as an outlier that would be better discarded to improve the normality of the distribution, but in fact she also displays important features of the variety of conditions for remote usage.

A standard procedure regarding the behavior of outliers 20 and 42 would be to discard them. But what happens when they are excluded from the sample? Figure 3 and figure 4 (below) show that the number of outliers increases without the two extremes. This is not surprising, since the elimination of the two extreme points changes the mean value and now, in relation to the mean, other scores become outliers. Figure 3 shows the *completion time in X*, depending on the setting. It is clearly visible that outliers occur in both settings—in the controlled laboratory environment and the natural environment—with a higher percentage of extreme outliers in the latter. It is apparent that there are too many extreme points (marked by asterisks) to suggest a normal distribution. It also becomes clear that the overall variable *test duration* masks the real test situation: rarely is one participant an outlier in several tasks.

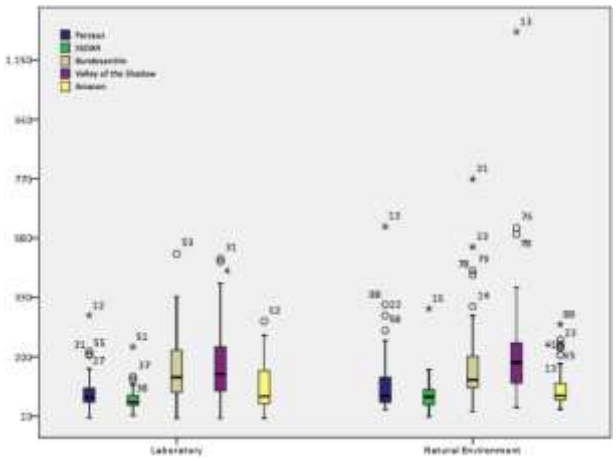


Figure 3. *Completion time in X* (y-axis) without participants 20 and 42.

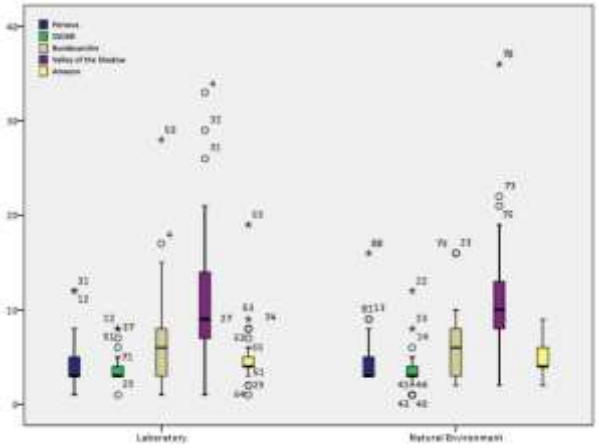


Figure 4. *Page views in X* (y-axis) without participants 20 and 42.

Participant 51 in the laboratory is an outlier on the task in the digital library *SSOAR*, but not in any other tasks. Participant 12 is only an outlier on the task in *Perseus*. This result is confirmed by Nielsen's findings: "The most seemingly obvious explanation for these outliers is simply that a few people are almost incompetent at using the Web, and they'll show up as slow outliers every time they test something. But this hypothesis is false. Once we recruit people for a study, we ask them to do multiple things, so we know how the slow outliers perform on several other tasks. In general, the same users who were extremely slow on some tasks were fast on other tasks" (Nielsen, 2006, online).

Figure 4 shows the same tasks in the five different digital libraries, but gives the variable *page views* on each task. Again, several outliers appear for both settings, with extreme points as well. However, the outliers in figure 3 are different from those in figure 4. For example, participant 21 was very slow, but did not need many page views to complete the tasks.

At this stage, it is necessary to describe other types of outliers that occur and to consider what makes these outliers become outliers. Is it higher variability on the distribution than expected, or is it only distraction that produces outliers? If the latter would it be enough to add a pause button to avoid outliers? Participant 13, also a female, offers a relevant example. She did the test at home, and for the whole test she needed close to an hour. Her average number of page views to complete a task were only 6.3 (compared to M = 5.9 page views). This means she was slow, but did not need many more clicks to complete the tasks than the average. She completed all German digital libraries (*SSOAR*, *Bundesarchiv* and *Amazon*) successfully, but abandoned both of the English language tasks, *Perseus* and *Valley of the Shadow*. She needed 617 seconds to finally abandon the task in *Perseus* (compared to the mean with 210 seconds) as well as many more page views (9 compared to the mean 4.3). This indicates that she had actively searched for a long time. The same was true for her behavior in *Valley of the Shadow*. Participant 13 said that she had a single application open during the test and that she looked at it three times, but she said that she did not feel distracted by it. She was also contacted once, but again felt little distraction. Compared to the other two outliers described above, she does not have a single obvious break that can be explained by a disturbance. At the end of the test, participants were asked to estimate their German and English skills and she rates her knowledge of English as very low: "I can only read it with trouble". A lack of language skills is unsurprisingly another factor that influences the distribution of time scores and page views.

A completely different type of outlier is illustrated by participant 4: a young man doing the test in the laboratory. With 1,869 seconds to complete the whole test, he is above the mean value of 1,327 seconds and he needed 10.5 page views (mean 5 page views). He also mentions technical problems with the digital library *Bundesarchiv*, which might explain the relative long time and page view score in that particular digital library. Nonetheless he was one of the few participants who were able to complete all tasks successfully. He says he is a PhD student. Based on the available information, participant 4 was neither distracted nor had problems completing the tasks. He was a very well intentioned participant who wanted to complete the test in the best possible way. Is participant 4 an outlier, because his behavior is too accurate to represent real retrieval behavior or does he only represent another end of the variability of data?

Figure 3 and 4 suggest that the data produce predominantly outliers on the upper end (that is, slow participants). A transformation into a standardized distribution by taking the logarithm of the time scores represents a standard method for solving the problem. A logarithmic transformation has "the consequence of bringing the tail involving slower latencies closer to the center of the distribution and making the mean a more accurate reflection of the central tendency of distribution" (Fazio, 1990, p. 85). The expectation is that the new logarithmic values will generate a standardized distribution. All statistical tests could then be performed—even those relating to the natural environment. To test this, a new set of boxplots was generated on the logarithmic values of the variables *completion time in X* (figure 3) and on the logarithmic values of the variables *page views in X* (figure 4), respectively figure 5 and figure 6 in logarithmic scale.

There are outliers in the natural environment (figure 5), but two outliers appear in the laboratory setting that have not been clearly visible before. These are "fast" outliers, ones at the lower end of the time scale. One could assume at this point that the new "fast" outliers behave in the same way as the "slow" outliers did: that is that they were outliers in one task and are not general outliers for all tasks. However, the boxplots of figure 5 and figure 6 show that the "fast" outliers follow another pattern. "Fast" outliers are outliers that appear in nearly every digital library and they are outliers both in the *completion time* and the *number of page views*. This means that these participants have neither spent much time on the individual tasks nor have they needed many clicks for the task.

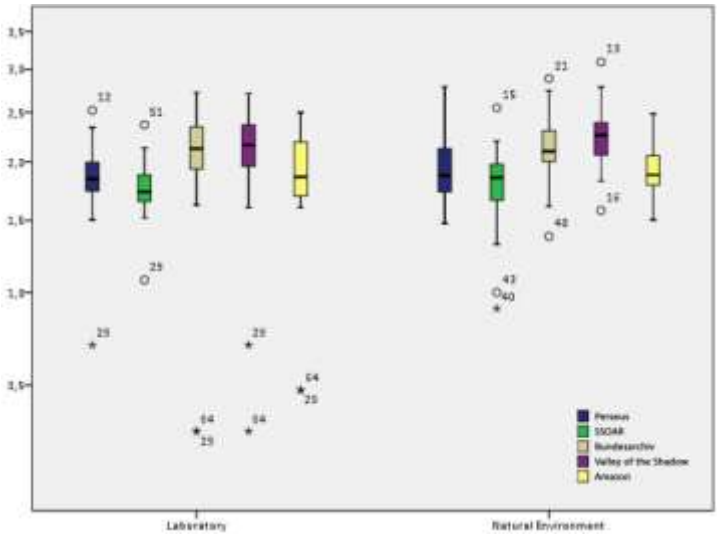


Figure 5. Logarithm: *Completion time in X* without participants 20 and 42.

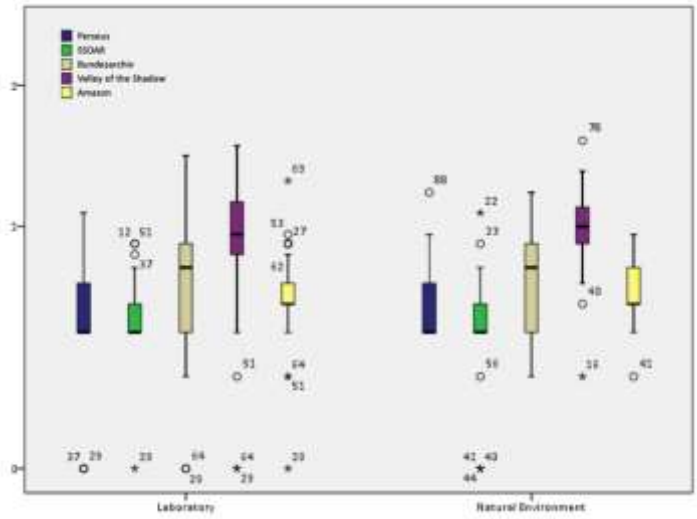


Figure 6. Logarithm: *Page views in X* without participants 20 and 42.

Participant 40, a male participant at home, needed only one page view in the digital library *SSOAR* and spent only 8 seconds on the task. His comment reveals that *SSOAR* did not load. This explains the one outlying score of this participant. Participant 41 and 43, both females doing the test at home on the same day, but some hours later, showed a similar behavior. Both report that *SSOAR* had problems loading and therefore appear as an outlier for that particular task.

In the laboratory setting two outliers appear several times in both boxplots: on the one for the five variables *completion time in X* and on the one for the variables *page views in X*. Participant 64 is a female whose native language was not German. She was the only one who said that she did not concentrate at all on the test (of course, this question came after the test and participants were assured that their answers had no influence on the reward). Apart from *SSOAR*, she failed all tasks. At the end, she abandoned the task in *Amazon*, maybe because she was no longer in the mood for tests. After her successful participation in *SSOAR*, she spent only between 2 and 5 seconds on each task; she finished all tasks in 119 seconds (M = 662 seconds) and took only 478 seconds on the whole test, including the post-questionnaire (M = 1327 seconds). It is interesting that she always clicked "task complete", even when she must have known that she had not found the right result.

Participant 29, a male, was the fastest participant in the test with 193 seconds on the whole test (M = 1327 seconds). Obviously, he could not have left the laboratory after 3 minutes, so it is to be assumed that after the test he continued on a personal task without being noticed by the recruiters in the laborato-

ry. He had an average page view rate of 1 click per task, which means he must have started the task and immediately clicked on "task complete". Like participant 64, he clicked on "task complete" even without having completed the task.

A last outlier in the laboratory was female participant 51. She follows a similar pattern to participant 64, but is smarter. She completed the task in *Valley of the Shadow* in 66 seconds and 2 page view. Anyone who has already used this particular digital library knows that this is simply impossible. A minimum of five page view was necessary to complete the task. However, her second page view was the requested page. It can only be speculated that she used a search engine like Google to find the right result.

It is interesting that all three participants, who in some sense cheated on the tasks, turned up in the laboratory under controlled circumstances. None of the participants in the natural environment obviously speeded through the test without actually doing it or using external help tools. This circumstance is remarkable, because it would have been much easier to cheat in the natural environment. It is also good news for tests in natural environment settings.

5. Categorization of outliers

Based on the previous analysis, outliers can be grouped into seven variations:

(1) The "speeder": This kind of outlier could theoretically appear in either a laboratory or a natural environment. Speeders are repetitive outliers on the variables *completion time in X* and the variable *page views in X*.
(2) The "scrupulous": This kind of outlier is mirrored by participants that are overly scrupulous in the test situation and try to accomplish tasks in the best possible way. Indicators are slower time scores than the mean and successful task completions for most tasks. This outlier can appear in either a laboratory or a natural environment.
(3) The "unlucky": This outlier, as identified by Nielsen (2006), mirrors a test situation in which the participant gets lost. It can appear in either a laboratory or a natural environment.
(4) The "multi-tasker": This kind of outlier appears in the natural environment. This outlier switches between tasks during the test time. The effect is an overall strong influence on the variable *test duration*. The number of page views of multi-tasker is close to the mean; and multi-tasker have a slightly less successful task performance.
(5) The "break-taker": This kind of outlier is marked by a strong external disturbance during the test, which requires the full attention of the participant. The break-taker can be identified by an exponentially slower time

score than the mean score. The outlier's page view score meets the mean. It appears in the natural environment.

(6) The "inculpable": This outlier is confronted with an external disturbing factor on which the participant has no influence. This can be, for example, a server problem. This outlier can only be identified by additional context information. It mostly appears in natural environment settings; participants in a laboratory could orally inform the researchers.

(7) The "handicapped": This outlier appears if participants do not have required competences for specific tasks, for example the language skills to search in English sites or a particular cultural or historical knowledge. This outlier can only be identified by adequate context information. This outlier can appear in either a laboratory or a natural environment.

6. Discussion

The previous analysis of outliers makes clear that outliers are not only caused by "bad luck". It also became obvious that distraction is an important (disturbing) factor in asynchronous remote tests. The analysis showed that not only distractions such as multitasking or a direct contact affected test data. The "inculpable" outlier meets with an external disturbing factor on which the participant has no influence. This can be, for example, a server outage. It is an inherent problem of online tests and context information is the only way to interpret that kind of outlying observation.

There are several human factors that have an influence on the data too, and are not caused by the natural environment. The "handicapped" outlier appears if participants do not have required competences for specific tasks, for example the language skills to search in English sites or a particular cultural or historical knowledge. While these factors also apply to non-remote test situations, they are all the more important in an asynchronous remote test setting in a natural environment, because researchers could not see when participants are having a problem, as they could in a laboratory setting.

The "scrupulous" outlier is a known phenomenon of test situations. It is mirrored by participants that are overly scrupulous in the test situation and try to accomplish tasks in the best possible way. The "speeder", on the other hand, explicitly manipulates a test situation. It might have been a chance circumstance that both types of outliers appeared only in the laboratory, and that these human factors seemed to play an insignificant role in natural environments. This sample was too small to establish that as true or not, but it raises the interesting possibility that online tests in natural environments could allow for a more adequate collection of real information behavior than laboratories.

This contribution suggests that outliers are not necessarily lost participants, but may, on the contrary, be an indication of normal events in the natural environment and thus an indispensable key to interpret information behavior. In this sense, outliers could even be better representations of real users than the statistically standard participant. They give evidence what has happened during a test situation: distractions, technical issues, frustrations – these are all part of user's life and have an influence on how people interact with information. Removing examples of everyday life behavior from the data set, because these do not fit statistical requirements, means reducing the external validity.

7. Conclusion

Data on time scores in asynchronous remote tests are not self-explanatory. A slow completion time must not necessarily mean that a participant had problems with a task. The interpretation of time scores or page views is not inherent to the measure itself. Researchers need context information to explain extreme values in information behavior studies. By simply claiming that these are outliers, the risk of losing important information on the user's behavior is immense.

This analysis can give no general advice what to do with the outliers. In the end, it depends on the research question that the data should answer. If the analysis consists of a statistical analysis there is probably no other possible way than to exclude the extreme points. This, however, requires a careful reconsideration of the goals of information behavior research: the primary aim should not be to run statistical analyses; instead, the aim should be to acquire new knowledge about user's information behavior. Acknowledging outliers and explaining the reason they became outliers would be a first step.

Outliers in natural environment settings do not necessarily indicate improper values. More likely, they indicate the existence of normal events in the environment, and should be welcomed as a key element in the interpretation of data. Outliers occur when some participants do not match the underlying assumed model of data distribution, but in the natural environment, the normal distribution is overlaid by processes that do not follow a normal distribution (for example interruptions or server problems during the test situation). These random processes cannot be completely excluded from remote test settings, since they are genuine parts of the natural environment.

Outliers do not only appear in information science studies. But the lack of reporting them appears to be genuine to the information science field. None of the fields with similar data – i.e. psychology, social sciences, and econom-

ics – show the same attitude towards non-reporting than information science. It does not increase the credibility of the field.

8. Acknowledgments

Thank you to Maria Yalpani, Pamela Aust, Ulrike Stöckel, Katja Metz, Nadine Messerschmidt, Kristin Reinhardt and Lars Gottschalk who helped with the recruiting process. I would also like to thank the software producers of *Loop11* for the free use of their product and Mareen Reichardt for her help on finding outliers in information science publications.

9. References /Literaturverzeichnis

Albert, W., Tullis, T., & Tedesco, D. (2010). Beyond the usability lab: Conducting large-scale online user experience studies. Amsterdam: Morgan Kaufmann/Elsevier.

Amanda H. Goodall. (2006). Should top universities be led by top researchers and are they? Journal of Documentation, 62(3), 388–411.

Amazon. Retrieved from http://www.amazon.de

Bates, M. J. (2010). Information Behavior. In M. J. Bates & M. N. Maack (Eds.), Encyclopedia of Library and Information Sciences (Vol. 33rd ed., pp. 2381–2391). New York: CRC Press.

Das Digitale Bildarchiv des Bundesarchivs. Retrieved from http://www.bild.bundesarchiv.de/

DigiZeitschriften: Das Deutsche Digitale Zeitschriftenarchiv. Retrieved from http://www.digizeitschriften.de/

Fazio, R. H. (1990). A practical guide to the use of response latency in social psychological research. In C. Hendrick & M. S. Clark (Eds.), Research methods in personality and social psychology (pp. 74–97). Thousand Oaks, CA, US: Sage Publications.

Gao, Y. (2004). Appeal of online computer games: a user perspective. The Electronic Library, 22(1), 74–78.

Greifeneder, E. (2012). Does it matter where we test? Online user studies in digital libraries in natural environments (Dissertation). Humboldt-Universität zu Berlin, Berlin. Retrieved from urn:nbn:de:kobv:11-100203293

Greifeneder, E. (2015). The effects of distraction on task completion scores in a natural environment test setting. JASIST, EarlyCite. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/asi.23537/abstract

Grupps, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. Technometrics, 11(1).

Heidar, M., Mohammad-Reza, D., Mohammad-Hossein, D., & Mohammad-Reza, A. (2013). Students' need for cognition affects their information seeking behavior. New Library World, 114(11/12), 542–549.

Huuskonen, S., & Vakkari, P. (2008). Students' search process and outcome in Medline in writing an essay for a class on evidence-based medicine. Journal of Documentation, 64(2), 287–303.

Ingwersen, P. (2007). Context in information interaction – revisited 2006. In T. Bothma & A. Kaniki (Eds.), ProLISSA 2006. Proceedings of the Fourth Biennial DISSAnet Conference (pp. 13–23).

Johnson, J. D. (2003). On contexts of information seeking. Information Processing & Management, 39(5), 735–760.

Joho, H., Jatowt, A., & Blanco, R. (2015). Temporal information searching behaviour and strategies. Information Processing & Management.

Kelly, B., & Peacock, I. (1999). How is my web community developing? Monitoring trends in web service provision. Journal of Documentation, 55(1), 82–95.

Kelly, D. (2006). Measuring online information seeking context, Part 1: Background and method. Journal of the American Society for Information Science and Technology, 57(14), 1729–1739.

Loop11: remote & online usability testing tool. Retrieved from http://www.loop11.com/

Miller, Z., Dickinson, B., Deitrick, W., Hu, W., & Wang, A. H. (2014). Twitter spammer detection using data stream clustering. Information Sciences, 260, 64–73.

Nielsen, J. (2006). Outliers and Luck in User Performance. Jakob Nielsen's Alertbox. Retrieved from http://www.useit.com/alertbox/outlier_performance.html

Perseus Digital Library. Retrieved from http://www.perseus.tufts.edu/hopper/

SSOAR: Social Science Open Access Repository. Retrieved from http://www.ssoar.info/

Tenopir, C., Volentine, R., & King, D. W. (2013). Social media and scholarly reading. Online Information Review, 37(2), 193–216.

The Valley of the Shadow: two communities in the American Civil War. Retrieved from http://valley.lib.virginia.edu/

Wijaya, D. T., & Bressan, S. (2006). Clustering web documents using co-citation, coupling, incoming, and outgoing hyperlinks: a comparative performance analysis of algorithms. International Journal of Web Information Systems, 2(2), 69–76.

Scales and Scores

An evaluation of methods to determine the intensity of subjective expressions

Josef Ruppenhofer, Jasper Brandes, Petra C. Steiner

Hildesheim University
Hildesheim, Germany
{ruppenho|brandesj|steinerp}@uni-hildesheim.de

Abstract

In this contribution, we present a survey of several methods that have been applied to the ordering of various types of subjective expressions (e.g. *good* < *great*), in particular adjectives and adverbs. Some of these methods use linguistic regularities that can be observed in large text corpora while others rely on external grounding in metadata, in particular the star ratings associated with product reviews. We discuss why these methods do not work uniformly across all types of expressions. We also present the first application of some of these methods to the intensity ordering of nouns (e.g. *moron* < *dummy*).

1. Introduction

While there is much interest in intensity ordering for application within sentiment analysis, the ability to assess the intensity associated of scalar expressions is a basic capability that NLP systems in general need to have. It is necessary for any NLP task that can be cast as a textual entailment problem, such as IR, Q&A, summarization etc. For instance, as illustrated by de Marneffe et al. (2010), when interpreting dialogue (A: *Was it good?* B: *It was ok / great / excellent*.), a yes/no question involving a gradable predicate may require understanding the entailment relations between that predicate and another contained in the answer.[9]

[9] The intensity ordering task within sentiment analysis can also be understood as an entailment problem, which is prefigured e.g. by Horn's (1976) discussion of conversational implicatures of scalar predicates.

Among gradable linguistic expressions, adjectives are the best-studied class. Various methods have been explored, some of which we will experiment with, namely phrasal patterns (Sheinman 2013; Melo and Bansal, 2013); using star ratings (Rill et al., 2012); extracting knowledge from lexical resources (Gatti and Guerini, 2012); and collostructional analysis (Ruppenhofer et al., 2014).

Less work has gone into the scalar properties of adverbs. Rill et al. (2012b) studied them indirectly in the context of ordering complex adjective phrases containing intensifying adverbs. In submitted work, we have experimented with extending and adapting the methods used for adjectival intensity ordering for use with adverbs.

As far as we know, only work in theoretical linguistics has analyzed intensity orderings among nouns (Morzycki, 2009).

2. Corpora and published ratings

For our experiments we use three corpora. The BNC and ukWaC are used to compute association measures and to mine for linguistic patterns. The Liu corpus of Amazon product reviews is used to project star ratings onto linguistic units. In addition, we evaluate Taboada et al.'s lexical resource as a source of intensity information.

Corpora	Tokens	Reference
Liu	~ 1.06 B	Jindal and Liu,
BNC	~ 0.1 B	Burnard, 2007
ukWaC	~ 2.25 B	Baroni et al., 2009
Lexicon	Entries	Reference
SoCaL	216 intensifying adv., 1549 nouns, 2827 adjectives	Taboada et al., 2011

Table 1. Corpora and published ratings

3. Scales

For the adverb ordering task, we use adjectives from 4 different semantic scales. These are shown in Table 2 together with their classification following Paradis (1997, 2001). The adverbs we used are shown below in Table 3, sorted into the classes defined by Paradis (1997). For the items used in the adjective ordering task, we refer to Ruppenhofer et al. (2014). The items of the noun ordering task are presented in Table 4.

Adjective	Scale	Pol.	Type
dumb	Intelligence	neg	scalar
smart	Intelligence	pos	scalar
brainless	Intelligence	neg	extreme
brainy	Intelligence	pos	extreme
bad	Quality	neg	scalar
good	Quality	pos	scalar
mediocre	Quality	neg	scalar
super	Quality	pos	extreme
cool	Temperature	neg	scalar
warm	Temperature	pos	scalar
frigid	Temperature	neg	extreme
hot	Temperature	pos	extreme
short	Duration	neg	scalar
long	Duration	pos	scalar
brief	Duration	neg	scalar
lengthy	Duration	pos	scalar

Table 2. Classification of adjectives used

Maximizer	Booster
absolutely	awfully
completely	extremely
perfectly	very
quite	highly
Moderator	Diminisher
quite	slightly
fairly	a little
pretty	somewhat
Approximator	Control
almost	no adverb

Table 4. Classification of adverbs used

Intelligence		Expertise	
positive	negative	positive	negative
Einstein, genius, brain, brainiac, superbrain, sage	blockhead, cretin, dimwit, doofus, fathead, fool, half-wit, idiot, imbecile, moron, nitwit	ace, adept, buff, champion, expert, guru, master, maven, pro, specialist, star, superstar, virtuoso, whiz	neophyte, newbie, novice

Table 3. Nouns used

4. Gold Standards

For all the items from the different scales, we elicited ratings using the online survey tool LimeSurvey.[10] We recruited our subjects from Amazon Mechanical Turk (AMT), specifying the following qualifications: US residency, a HIT-approval rate of at least 97%, and 500 prior completed HITs.

[10] www.limesurvey.org

The surveys typically used several parallel surveys, each eliciting data for subsets of our items, to be completed by non-overlapping sets of participants, which we controlled by checking AMT worker IDs. In each survey, participants were first asked for metadata such as age, residency, native language etc. Each survey used pairs of main and distractor block and was concluded at the end by a block in which feedback / comments on the survey was invited. All items were rated individually. The blocks and the items in the blocks were randomized so as to minimize bias. Participants were asked to use a horizontal slider, dragging it in the desired direction, representing polarity, and releasing the mouse at the desired intensity, ranging from −100 to +100.

5. Methods

5.1 Horn patterns

Horn (1976) put forth a set of pattern-based diagnostics for acquiring information about the relative intensity of linguistic items that express different degrees of some underlying property. The complete set of seven diagnostics is shown in Table 5.

For all patterns, the item in the Y slot needs to be stronger than that in the X slot. The two slots can be filled by different types of expressions such as adjectives, nouns, and adjectives, as shown by the following examples.

(1) It's not just *entertaining* but *hilarious*. (adjectives)
(2) Peter's a *dummy,* or even an *idiot*. (nouns)
(3) This is *very* good, if not *extremely* good. (adverbs, with adjective held constant)

Based on the frequencies with which different items of a specific type occur in the X and Y slots, we can induce a ranking of these items.

X (,) and in fact Y	not only X(,) but Y
X (,) or even Y	not X, let alone Y
X (,) if not Y	not Y, not even X
be X (,) but not Y	

Table 2 Horn patterns

5.2 MeanStar

Another corpus-based method we evaluate employs mean star ratings derived from product reviews, as described by Rill et al. (2012b). Note that this method uses no linguistic properties intrinsic in the text. Instead, it derives intensity for items in the review texts from the numeric star ratings that reviewers (manually) assign to products. Generalizing the approach of Rill et al. (2012b) to any kind of simple or complex unit, we define the intensity score for a unit as the weighted mean of the star ratings

$$SR_i = \frac{\sum_{j=1}^{n} \quad S_j^i}{n}$$

where *i* designates a distinct unit, *j* is the *j*-th occurrence of the unit, S^i_j is the star rating of *i* in *j*, and *n* is the total of observed instances of unit *i*.

5.3 Collexeme analysis (Collex)

Collexeme analysis (Gries and Stefanowitsch, 2004) exploits the association strength between linguistic units and the constructions that they can occur in. For instance, in the case of adjectives, one assumes that adjectives with different types of intensities co-occur with different types of adverbial modifiers. End-of-scale modifiers such as *extremely* or *absolutely* target adjectives with a partially or fully closed scale (in the sense of Kennedy and McNally (2005)), such as *brilliant* or *outstanding*, which occupy extreme positions on the intensity scale. 'Normal' degree modifiers such as *very* or *rather* target adjectives with an open scale structure, such as *good* or *decent*, which occupy non-extreme positions.

To determine a linguistic unit's preference for one of two constructions, the Fisher exact test (Pedersen, 1996) is used. It makes no distributional assumptions and does not require a minimum sample size. The direction in which observed values differ from expected ones indicates a preference for one construction over the other and the p-values are taken as a measure of the preference strength.

In the case of adjectives, our hypothesis is that e.g. an adjective A with greater preference for the end-of-scale construction than adjective B has a greater inherent intensity than B.

Note that Collex produces two rankings, one representing the degree of attraction to one of the constructions. To obtain a global intensity ordering, they need to be combined. In the case of ordering adjectives, the positive/negative adjectives being attracted to the extreme modification construc-

tion were put at the top/bottom of the ranking. The set of adjectives that prefer the normal modification construction are placed between the extreme positive and negative sets. Here, the positive/negative adjective least attracted to the normal construction immediately adjoins the positive/negative adjective least attracted to the extreme construction. Adjectives that have no preference for either construction are finally inserted in between the positive and negative adjectives attracted to the normal construction.

For adverbs, we consider the adjective-adverb nexus in the opposite direction: the adverbs are the units to score and classes of adjectives define the different constructions. For nouns, we can proceed in simple analogy to the case of adjectives, except that the modifiers of nouns are adjectives such as *high* or *utter* rather than adverbs such as *highly* or *utterly*.

6. Experiments

6.1 Adjectives

In earlier work (Ruppenhofer et al., 2014), we compared the performance of our methods on both subjective adjectives as well as objectives ones. We found Collex to give good performance for both types of adjectives. While de Melo and Bansal (2013) report very good results using Horn patterns, we prefer the use of Collex because it does not need web-scale data (Google 5-grams), working even on 'smaller' corpora such as the BNC, and is computationally simpler than the sophisticated interpolation approach applied by those authors. The MeanStar method was slightly better than Collex for subjective adjectives but very low-performing for objective ones. Of the lexical resources we considered, SoCAL had the best results. However, SoCAL has coverage gaps for objective adjectives.

6.2 Adverbs

Horn patterns cannot be used for adverbs, at least not currently. In the ukWaC, there are very few instances of Horn's 7 patterns that have two different adverbs but the same adjective in the X and Y slots. The frequency of relevant adverb-adjective instances is in fact significantly lower than that of simple adjective instances. On web-scale data, this approach might still become feasible. However, it is currently not feasible because for the smallest pattern to host two adverb-adjective pairs in the X and Y slots, one would already need 6-grams, whereas only 5-grams are available.

The collostructional approach also did not perform well, counter to our initial hopes and expectations. Using the same ranking strategy that Ruppenhofer et al. (2014) employed for adjectives (cf. section 5.3) but with adjectives and adverbs switching roles, produced very low correlation results below 0.2. In hindsight, we believe that this is due to a significant asymmetry between adverbs and adjectives. Among adjectives, the extreme and scalar subgroups are the largest and they tend to be well separated: scalar adjectives tend not to have intensities as great or greater than extreme adjectives. Adverbs are different. First, the gold standard data suggest that adverbs in distinct classes do not have separate bands of scaling effect. For instance, of all adverbs, *extremely*, a booster, has the highest scaling effect, at least matching if not outdoing maximizers such as *utterly* and *absolutely*. And while moderators and diminishers are separated pretty well in the human ratings, the approximator *almost* is sandwiched among the diminishers. The Collex approach is not set up to handle this constellation well since, as shown in Figure 1, it expects to find maximizers and approximators to be most drawn to limit and extreme adjectives, and boosters, moderators and diminishers to be attracted by scalar adjectives. Thus, maximizers and approximators should have similar and consistently higher scaling effects than the other types of intensifiers. Reality fails to comply with this assumption and accordingly we obtain poor results. Collex is thus a one-way strategy: it can rank adjectives based on adverbs but not the other way around.

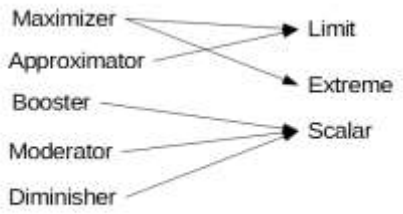


Figure 1. Adverb-adjective interaction

SoCaL's intensifier lexicon has good coverage for our adverbs and provides intensity scores for all of the items. A ranking of our adverbs as obtained by these intensity scores produces near-perfect correlations with our gold standard (0.97). Nonetheless, a drawback of relying on lexical resources for information on the scaling effect of adverbs is that, while there is class of frequently used and highly grammaticized ones, such as the ones we considered here, there is a much larger, fluid set of less grammaticized adverbs (e.g. *preternaturally*) that lexicons will be unlikely to ever fully cover. For these cases having a corpus-based method is necessary.

We finally consider the MeanStar approach. That approach should not in theory be useable directly with adverbs by themselves since they do not have an inherent intensity like adjectives or nouns do but instead act upon the intensity of the predicates they modify. Nevertheless, as a baseline measure we tried the brute force approach of projecting star ratings onto adverbs regardless of the adjectives. The results were better than what we obtained with the collostructional approach: a correlation of 0.283 when using all instances of adverbs found anywhere, and a correlation of 0.446 when only taking into account instances occurring in review titles. This difference between review bodies and titles has been observed before by Rill et al. (2012b) and stems from the fact that titles tend to more straightforwardly match the tenor of the star rating, while review bodies may offer discussions of pros and cons that do not align as cleanly with the star rating given.

Intuitively, if we want to improve upon the adverb-only baseline, we had best taken into account the adjectives being modified by the adverbs. Ideally, we would find every adverb we want to rank used in combination with every adjective that we want to work with. On that basis, we could learn to 'factor' out the effect of the adverb by comparing the scores of adverb-adjective combinations involving the same adjective, to each other and to the score of the unmodified adjective. However, here too, we run up against the actual distribution, which is not as we would like it to be. As the log-log plot in Figure 2 shows, there are many adjectives that occur with a few adverbs and few adjectives that occur with many. We therefore do not find all the combinations that we would need to have so that we could produce per-adjective rankings of the adverbs, which we could then combine into a global ranking of the adverbs.

This distributional fact doomed the first method that we experimented with, which tried to integrate the relative intensity differences between adverb-adjective combinations and other combinations and the simple adjectives, by observing which combinations tend to have greater scores than others. Technically, this was a use of the Borda count method from Voting theory, where voters rank some number of candidates in their order of preference. The adjectives can be thought of as the 'voters' on the ranking of the adverbs. However, this approach performed badly because with our data, we fail to satisfy a core assumption of Borda count, namely that candidates not voted for (i.e. unobserved adverb-adjective combinations) should be ranked lower than any candidates voted for (i.e. observed adverb-adjective combinations).

Actually, even the combinations per adjective that we do find are somewhat deceptive. As shown by the work of Desagulier (2014), adjectives may prefer to co-occur for instance with different moderators depending on the specific word sense involved. As an illustration, consider that in the ukWaC corpus

the combination *pretty cool* is almost 100% associated with the desirability sense of *cool* found in e.g. *cool idea*! By contrast, the combination *fairly cool* is almost exclusively used in the temperature sense found e.g. in *cool weather*. Any corpus-based method must thus make the bet that the most frequent readings of most adjectives will nevertheless belong to the same adjective type in the sense of Paradis and can thus be conflated together.

Accepting that one needs to deal with lemma-level data and pursuing an approach that tries to capture an adverb's scaling effect against the simple adjective, there remains the problem of how to conceive of that scaling effect. In the context of research on review mining, Liu and Seneff (2009) model it as the difference between the intensity of an adverb-adjective combination and the intensity of single adjective. They did not, however, evaluate their model directly against human ratings as a gold standard but only extrinsically as part of an automatic system. It is therefore not clear how well their model of adverb intensity works.

In work of our own that is currently under review, we have pursued a different approach of conceiving of the scaling effect. Basically, we try to capture the relative scaling effect rather than the absolute distance. For example, if we measure the difference between *absolutely good* and simple *good*, and between *absolutely perfect* and simple *perfect*, then on the Liu and Seneff approach, *absolutely* will seem to have a weaker effect on the adjective *perfect* than on the adjective *good* because *perfect* has a higher intensity to start with. Our approach instead asks: how far does the adverb move the adjective's intensity towards the end of the scale, relative to the available distance to be covered? On that approach, the scaling effect of *absolutely* will seem substantial even when applied to *perfect*. We obtained very good results for this method but comparing it to the Liu and Seneff approach remains for future work.

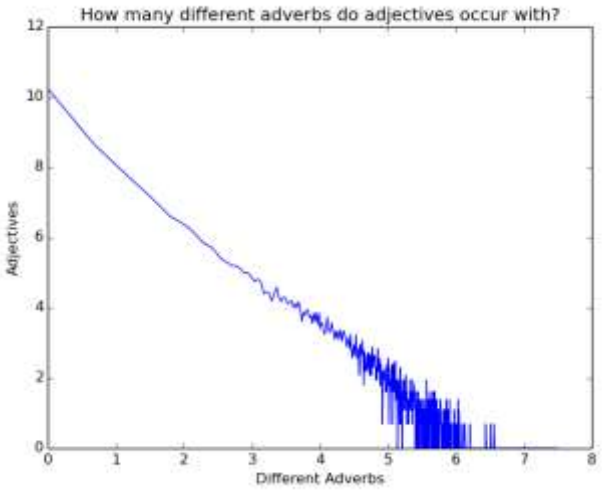


Figure 2. Adverb-adjective co-occurrence in the ukWaC

6.3 Nouns

As for the adjectives and adverbs, for the nouns we also come across severe coverage issues for Horn's patterns. For none of the 7 patterns, we find instances where there are two different nouns from one of the two examined scales. This even holds true if we loosen the constraint and allow up to three additional tokens in between the determiner and the noun. Thus, we currently see no way of using Horn's patterns for nouns.

SoCaL's coverage for nouns is poorest across the three types of expressions investigated in this study: there are intensity scores for only 4 of the 17 intelligence nouns and only for 5 of the 17 expertise nouns. As such, at least for our two scales, SoCaL cannot be used for the intensity ordering of nouns referring to the same scale.

MeanStar produces low positive correlations (0.2) for the intelligence nouns and medium positive correlations (0.51) for the expertise nouns when performed on the review titles. While these results are not good, for the intelligence nouns they can be attributed to the low frequencies of these nouns in the review titles. Coverage, however, is quite good with 15 intelligence and 16 expertise nouns occurring in the review titles.

Finally, we report on the results for Collex. We followed the same approach as for the adjectives, only that for the nouns, we replaced the adverbs with adjectives (i.e. *high* instead of *highly* and *utter* instead of *utterly*). We distin-

guish between two constructions a noun can occur in: modification by 'end-of-scale' adjectives such as *utter* or *complete* or by 'normal' adjectives such as *big* or *slight*. We assume that nouns which are more attracted to 'end-of-scale' adjectives have a higher inherent intensity than nouns that are more attracted to 'normal' adjectives. The ranking approach put forward in Ruppenhofer et al. (2014), yields medium correlations (0.58) for the expertise nouns and high correlations (0.89) for the intelligence nouns.

7. Conclusion

In this paper, we presented a discussion of methods for determining the intensity of subjective expressions. We focused on different semantic scales of English adverbs, adjectives, and nouns. In the case of adjectives and nouns, we have examined both subjective and objective scales.

None of the presented methods works universally well for all considered types of expressions. While Horn's patterns (e.g. 'X or even Y'), one of the two linguistically grounded methods, seem promising, severe coverage issues make this approach currently unusable. The other linguistically motivated method, Collex, works very well for adjectives and quite well for nouns, while for adverbs correlations with a human gold standard are very low. MeanStar produces good correlations for the adjectives and low to medium correlations for adverbs and nouns. Note that the MeanStar approach is dependent on (manually assigned) metadata from a large review corpus which may not be available for all languages. This is also the case for lexical resources which assign intensity ratings to lexical items. SoCaL, a much-cited subjectivity lexicon, fares well for the adjectives and adverbs but has very low coverage for nouns. The pursuit of corpus-based methods is thus necessary for reasons of coverage. Potentially, it is also interesting for building customized intensity ratings, for instance, for the American versus the British variety of English, but also for specific application contexts such as product review mining.

8. References/Literaturverzeichnis

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetti. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. Language Resources and Evaluation, 43(3):209-226.

Burnard, Lou. (2007). Reference Guide for the British National Corpus, Research Technologies Service at Oxford University Computing Services, Oxford, UK.

Desagulier, Guillaume. (2014). Corpus Methods for Semantics, chapter Visualizing distances in a set of near-synonyms, pages 145-178. John Benjamins Publishing Company, Amsterdam, Philadelphia.

de Marneffe, Marie-Catherine, Christopher D. Manning, and Christopher Potts. (2010). Was it good? It was provocative. Learning the meaning of scalar adjectives. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. ACL '10, pages 167-176, Stroudsberg, PA, USA. Association for Computational Linguistics.

de Melo, Gerard G. and Mohit Bansal. (2013). Good, Great, Excellent: Global Inference of Semantic Intensities. In: Transactions of the Association for Computational Linguistics 1:279-290

Gatti, Lorenzo & Marco Guerini (2012). Assessing Sentiment Strength in Words Prior Polarities. Proceedings of the International Conference on Computational Linguistics (COLING). 2012, 361-370.

Gries, Stefan Th. and Anatol Stefanowitsch. (2004). Extending collostructional analysis: a corpus-based perspective on alternations. International Journal of Corpus Linguistics, 9(1):97-129.

Horn, Laurence Robert. (1976). On the Semantic Properties of Logical Operators in English. Indiana University Linguistics Club.

Jindal, Nitin and Bing Liu. (2008). Opinion Spam and Analysis. Proceedings of the international conference on Web search and web data mining (WSDM), pages 219-230.

Kennedy, Christopher and Louise McNally. (2005). Scale Structure, Degree Modification, and the Semantics of Gradable Predicates. Language, 81(2):345-338.

Liu, Jingjing and Stephanie Seneff. (2009). Review Sentiment Scoring via a Parse-and-Paraphrase Paradigm. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 161-169

Morzycki, Marcin. (2009). Degree modification of gradable nouns: size adjectives and adnominal degree morphemes. Natural Language Semantics 17(2):175-203.

Paradis, Carita. (1997). Degree modifiers of adjectives in spoken British English. Lund University Press.

Paradis, Carita. (2001). Adjectives and boundedness. Cognitive Linguistics, (12):47-65.

Pedersen, Ted. (1996). Fishing for Exactness. Proceedings of the South-Central SAS Users Group Conference.

Rill, Sven, Sven Adolph, Johannes Drescher, Dirk Reinel, Jörg Scheidt, Oliver Schütz, Florian Wogenstein, Roberto V. Zicari, and Nikolaos Korfiatis. (2012a). A phrase-based opinion list for the German language. Proceedings of KONVENS, 305-313.

Rill, Sven, Johannes Drescher, Dirk Reinel, Jörg Scheidt, Oliver Schuetz, Florian Wogenstein, and Daniel Simon. (2012b). A Generic Approach to Generate Opinion Lists of Phrases for Opinion Mining Applications. Proceedings of the KDD-Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM).

Ruppenhofer, Josef, Michael Wiegand, and Jasper Brandes. (2014). Comparing methods for deriving intensity scores for adjectives. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers, pages 117-122, Gothenburg, Sweden, April. Association for Computational Linguistics.

Sheinman, Vera & Takenobu Tokunaga. (2009). AdjScales: Differentiating between Similar Adjectives for Language Learners. CSEDU 1:229-235.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. (2011). Lexicon-Based Methods for Sentiment Analysis. In: Computational Linguistics 37 (2):267-307.

Patent Analysis and Patent Clustering for Technology Trend Mining

Noushin Fadaei[1]; Thomas Mandl[1]; Michael Schwantner[2]; Mustafa Sofean[2]; Julia M. Struß[1]; Katrin Werner[1]; Christa Womser-Hacker[1]

[1]University of Hildesheim
Marienburger Platz 22
31141 Hildesheim
fadaei@uni-hildesheim.de

[2]FIZ Karlsruhe
Hermann-von-Helmholtz Platz 1
76344 Eggenstein-Leopoldshafen
mustafa.sofean@fiz-karlsruhe.de

Abstract

Recognizing the emergence of new trends in technology comprises an invaluable competitive edge for industrial companies. A natural candidate to extract this information is given by patent databases, which by their very nature collect recent activities of major research and development departments. In order to identify trends within a patent collection the available data mostly is considered to be clustered in particular with respect to time. Considering patents as textual data, this paper analyzes the performance of K-means as the most common text mining algorithm for clustering patents, discusses challenges arising from the particular structure of the patents and suggests approaches to cope with these difficulties. In particular, we develop an evaluation procedure that assesses the quality of cluster solutions based on freely available queries to retrieve gold standards without the need for additional juror judgments.

1. Introduction

As for planning the research strategies, one of the vital and primary challenges that research institutes as well as industry are facing is recognizing trends. Patent databases as fast growing information resources represent a target database at this level of preparation. According to the annual report of the European Patent Office the patent filings in 2013 show an increase of 2.9% compared to 2012 and another increase of 3.1% is observed in 2014 compared to the number of patents filed the previous year (European Patent Office, 2014). Such statistical reports represent the growing competition in research projects which emphasizes trend detection at early stages of formation.

Interviewing scientists and information professionals, Struß et al. (2014) have collected qualitative information on characteristics of trends. Their study collects factors usually considered for trend detection by professionals such as the number of publications in one particular area, the appearance of new IPC classes or frequent co-occurrence of IPC classes. Consequently, to investigate technical trends an automatic system initially needs to identify classes of patents in terms of their technological objectives. Further analysis on these classes will help to predict trends among them.

In this paper we investigate the classification of patent data using K-means as it is one of the most common textual clustering algorithms. We show how this algorithm is influenced by the diversity of the content of patents and to which extent it is beneficial for the trend mining purposes.

The rest of the paper is organized as follows. The next section presents the key studies that are related to our work. Sections 3 and 4 discuss the steps our approach takes in order to cluster patent data. In section 5 we observe some first experimental results and analyze them. We have concluded the study and suggested future works in section 6.

2. Related Work

High competition in various research areas and the necessity of detecting technological trends at initial stages has recently motivated several studies on patent analysis (Struß et al. (2014), Yoon et al. (2011), Choi et al. (2011), Chang et al. (2010)). These works take advantage of diverse aspects of a patent's structure; however the focus is mostly either on bibliographic information or the content of patents. Considering patents as textual data, a range of text mining techniques are exploited in order to provide feature extractions upon which patent clusters or networks can be built.

A method proposed by Streibel (2008) combines both non-semantic and semantic sides of the patent text as target features. The former is gained by extracting adjective-noun pairs which may explain the sentiment of the patent while the latter needs experts to build comprehensive knowledge bases using Extreme Tagging System. Our study supports the keywords selection that is used in the foregoing non-semantic text analysis.

Other approaches rely on relation extraction using dependency parsers. Yoon et al. (2011) proposes a methodology that strives to build a network on the extracted function and property relations; similarly Choi et al. (2011) shed light on Subject-Action-Object relations of the patent full text. Both methods generate co-occurrence matrices out of the extracted relation from which the

final networks are constructed. The second approach names clustering as a further step in the selection of the patent data which happens before feature extraction yet its causality has not been elaborated.

Chang et al. (2010) survey the relationship of patents in an entire network as well as at cluster level. The network construction is based on the Euclidean distance of the feature vectors reflecting relevant patent keywords. To obtain more information about the network some indexes have been introduced to depict the centrality, progress or connectivity in that network. This study points out that the aforementioned measures are applicable at cluster level. To this end, the given data is clustered using multivariate analysis which combines hierarchical and nonhierarchical clustering methods; however they have not provided any more information on the selected algorithms nor the evaluation of the procedure. In conclusion we see that text mining techniques in general and clustering methods in particular can be valuable for detecting similarities or distinguishing differences between patents. This can be seen as an important prerequisite for the identification of trends in technologies.

3. K-means Patent Clustering

To monitor technological deviations inside a patent pool, one needs to determine when and to which part the preferred types of changes occur; patents provide us with some temporal information such as the publication date, nonetheless the information on a certain technology needs to be retrieved. Having thematic clusters of patents associated with their publication dates reduce the problem to cluster analysis over time.

Considering patents as textual data, we develop and evaluate k-means as the most common textual clustering algorithm. The procedure comprises the patent collection, the patent text preprocessing, the organization of vectors for each patent and k-means patent clustering. Each of these steps is further explained in the following subsections.

3.1 Patent collection

The first step to build clusters of patents is to organize them with regard to the technology they are describing. There are two components designed to obtain patents within the specific area name patent fetcher and patent keeper. The patent fetcher uses a search query to retrieve the related patents. The search strategy focuses on keywords and IPC-classes of English patents. Since the final purpose of clustering is trend identification the date range should also be a search criterion. As a result we obtain a list of documents

representing with their various fields of information including ID, description, publication date, inventors and IPC. The patent keeper is responsible to handle and store the patent information which was returned by the patent fetcher. It also comprises some functions which are required for updating.

3.2 Patent text preprocessing

Having the target patent texts, a range of Natural Language Processing (NLP) techniques are performed for extracting and reducing features. First of all, the sentences should be identified so the rest of the NLP methods can be applied to them. Once the tokenization of the sentences is completed, the tokens are tagged by their Part-of-Speech (PoS), using the Stanford parser. To attain keywords of the text, stop words like articles, prepositions and frequent words are eliminated from the text. We also utilize the document frequency to identify highly frequent words which can be treated as stop words, too. Stemming is another task which reduces different forms of words into their stems; this helps with counting the words which are semantically the same while they appear in various forms.

3.3 Organizing vectors for patents

Before the preprocessed text is ready to be clustered, it still needs to be converted into a format which can be understood by the k-means clustering algorithm. In information retrieval one useful scheme to associate weights to the keywords is TF-IDF. TF stands for term frequency in a document and measures the significance of a word within a document or its relevance to a query word. However, it is not enough to consider the overall score of a term; IDF stands for inverse term frequency; the higher this number the less frequent is the keyword within the document collection and thus is more likely to be informative and relevant. In our approach noun phrases (NPs) are selected as features. These features are then scored by TF-IDF and vectors whose dimensions are reflected by the mentioned scores represent the corresponding document.

3.4 Patent clustering

The standard K-means clustering receives a natural number K as input and generates K centers in between the distributed data points. Each data point is then assigned to one of the centers based on its distance to it. In our study, the vectors generated for patents compete for the centers based on the cosine similarity.

Table 1 depicts K-means produces promising results when the number of K is correctly determined in advance. However this requires sufficient knowledge about the data to be clustered. In the patent domain this task can only be completed with the help of experts. X-means was inquired to estimate the number of clusters for patent clustering task.

X-means is a method which tries to detect the optimal value of K in K-means clustering on the fly (Pelleg and Moore, 2000). Starting from an initial K it performs an additional K-means step inside each cluster by splitting each single center into two (K=2). Comparing the Bayesian Information Criterion (BIC) of the new cluster solution and the parent cluster (K=1) the final decision on allowing new clusters is made. This method did not result in generating the desired K for our data distribution as it tended to return the minimum given threshold. So for an initial estimation we assumed the square root of half of the given documents to determine K.

Topics	#Docs	Correct Docs in classes(k-Means with K=6)
Cryptography	50	50
Information Retrieval	50	50
Ink Jet Printer	50	49
Solar Energy	50	50
Touch Screen	50	45
Treatment of Cancer	50	46
-		10
Total	300	

Table1. Results of K-means clustering on patents belonging to non-similar topics

4. Experiment

In this section we describe our experimental setup to obtain several classes of patents. We used freely available queries that are provided by the World Intellectual Property Organization (WIPO) to form a gold standard. The result sets were gained through the EPFULL repository.

There are a number of technological reports available by WIPO under the universal title of “public health/ life sciences”. These reports usually cover

two types of queries; one results in the main class (usually shares the title with the topic that is reported) and one is breaking the main query into pieces; thus produces subclasses. Some reports categorize the foregoing subclasses and provide fewer yet more general subclasses.

Depending on the authors of the reports, queries are described in different languages and formats. Therefore the queries had to be adjusted to match Json data type. Using Elasticsearch's API, we collected data through the European Patent Fulltext (EPFULL) database and obtained the main class and subsequently the corresponding subclasses. For instance under the topic "3D printing", 163 patents were retrieved, out of which 102 patents were covered by subclasses from which 55 hits belong to technologies, 189 hits belong to materials and 38 hits belong to applications. Patents may share different subclasses and some available patents in EPFULL may not be covered by any subclass.

Retrieved patents of each main class are then passed to the K-means clustering system. Results and further analysis is discussed in the next section.

5. Results

In this section the discussion will point to the performance of the K-means clustering system given the patent data fetched from EPFULL, we provide real-world examples for our claims and suggest approaches that may cope with the observed downsides.

5.1 Evaluation criterion

For the evaluation task we consider the patents which appear in the gold standard and measure them with external validations which make use of the gold standard; they can cover a range of highly used measures such as purity, mutual information and F-measure. All the mentioned measures are considered in our study; however as all the patents will be assigned to particular clusters a retrieval measure such as the F-measure does not seem to be very manageable. Also, to consider the F-measure for each cluster the corresponding class is required. This would demand matching the number of clusters and the real-life classes.

5.2 Experimental Results

The retrieved patents described in the previous section are assigned to the clusters by the K-means algorithm. The contingency table of the topic "3D printing" in technology criterion is displayed in table 2. Regarding the total number of patents, we determined K=10 so that the table shows the resulting clusters in the rows. The columns display the different technologies that are specialized in the WIPO report. The field values are the number of patents in the EPFULL associated with the corresponding category. One apparent observation is the fact that once a row or cluster contains a high number of patents, the rest is becoming quite sparse.

The F-measure, the precision and the recall reflect the same issue. The measures purity and mutual information are described in table 3. The more the objects inside clusters are similar to each other, the higher is the purity measure. Generally we observe no purity better than 50%. The mutual information describes the accuracy of assigning patents of subclasses to clusters remains below 0.22.

5.3 Error analysis

The first row of table 2 indicates that the highest proportion of the patents is assigned to cluster C1. Generally the first three rows of the table consist of 31 patents whereas the last three rows (C8 to C10) contain only 12 patents. In K-means algorithm, once a data point is taken by one cluster, it cannot occur in any other cluster. This means once a patent is sorted into a specific cluster, it is not possible to resort it into another cluster while patents by nature share different aspects of technologies, materials and may have several applications. It drops the accuracy of the K-means drastically, especially if that criterion is very probable for allowing intersections of subclasses.

Under the criterion material patents in one subclass called "ABS plastic" are shared with 11 (over 20% of) subclasses on average while under application criterion we observe patents in "mechanical" subclass are shared with 1.8 (18%) other subclasses on average. The mutual information measures for these are respectively 0.11 and 0.22.

The relatively low purity measures in Table 3 can be explained by relatively high percentage of nonsimilar documents in each cluster which might belong to another non-defined cluster. Though this circumstance reveals that K-means may not work properly on patent data, which is particularly due to lack of knowledge on the possible number of clusters, it is also in line with the problem of unique assignments in K-means; if the clustering algorithm

would allow for the assignment of patents into multiple clusters, this would positively influence the purity.

To have an efficient evaluation we need the equal number of clusters and subclasses. This again emphasizes the problem of finding an optimal K in K-means algorithm. As in many evaluations we need to map the clusters and the subclasses to be able to compute accuracy.

6. Conclusion and Future work

In this paper we surveyed the problem of patent clustering with K-means as the most common text classification algorithm. Even though it shows promising results at the level of barely related topics or at least fields with few overlaps, it still shows significant drawbacks regarding categorizing texts thematically under the same criterion. One difficulty is to estimate an optimal number of clusters, i.e. K. Algorithms like X-means and bottom-lines don't work desirably in this domain; yet other methods required to be investigated for the matter.

One other struggle for improving K-means clustering is to explore methods which update the methodology in a way that it allows for overlaps between clusters. As in the case of patents, they naturally target different technologies, materials or applications and share information between different classes.

In future studies keys to aforementioned difficulties will be explored and the output clusters will be analyzed over time so we can also estimate the use of the optimal patent clustering in the task of trend mining.

Technologies	Electron Beam Melting	Fused Deposition Modeling	Inkjet Deposition	Laminated Object Manufactur-ing	Laser Engineered Net Shaping	Laser Metal Forming	Cladding	Multiphoton Lithography	Photolithography	Robocasting	Selective Laser Melting	Computer Numerical Control (CNC)	Selective Laser Sintering	Solid Ground Curing	Spin Casting	Stereolithography	Direct Metal Deposition
#	2	2	1	6	1	1	2	1	2	1	8	3	10	1	1	10	3
C1	1	0	0	1	1	1	0	0	0	0	5	1	5	0	0	2	1
C2	0	0	1	1	0	0	0	0	0	0	1	1	2	0	0	1	1
C3	0	0	0	0	0	0	0	1	2	0	1	0	0	0	1	0	0
C4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C5	0	0	0	1	0	0	1	0	0	0	0	1	1	0	0	2	0
C6	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
C7	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1
C8	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	3	0
C9	0	2	0	1	0	0	0	0	0	0	0	0	1	1	0	1	0
C10	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0

Table2. Contingency table of the topic “3D printing” in technology criterion

	Technology	Material	Application
Purity	0.35	0.15	0.50
Mutual Information	0.17	0.11	0.22

Table3. Purity and Mutual Information measure of different criterions of “3D printing”

7. References/Literaturverzeichnis

European Patent Office. Annual Report 2014: Statistics and trends: Total European patent filings, 2013. Online available at: http://www.epo.org/about-us/annual-reports-statistics/annual-report/2014/statistics/patent-filings.html

Chang P.-L., Wu C.-C., and Leu H.-J. (2010). Using Patent Analyses to Monitor the Technological Trends in an Emerging Field of Technology: a Case of Carbon Nanotube Field Emission Display. *Scientometrics*, 82(1):5-19.

Choi S., Yoon J., Kim K., Lee J. Y., and Kim C.-H. (2011). SAO Network Analysis of Patents for Technology Trends Identi_cation: a Case Study of Polymer Electrolyte Membrane Technology in Proton Exchange Membrane Fuel Cells. *Scientometrics*, 88(3):863-883.

Pelleg D. and Moore. A. W. (2000). X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning* (ICML '00), Pat Langley (Ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 727-734.

Olga Streibel (June 2008). Trend Mining with Semantic-Based Learning European Semantic Web Conference ESWC2008, PhD Symposium, Teneriffe, Spain.

Struß, J.; Mandl, T.; Schwantner, M. and Womser-Hacker, C. (2014). Understanding Trends in the Patent Domain. User Perceptions on Trends and Trend Related Concepts. In : *Proceedings of the First International Workshop on Patent Mining and Its Applications*. At KONVENS'14. IPaMin. Hildesheim, (October 7, 2014).

Yoon J., Choi S., and Kim K. (2011). Invention Property-function Network Analysis of Patents: a Case of Silicon-based Thin Film Solar Cells. *Scientometrics*, 86(3):687-703.

Session 3
E-Learning und Informationsmarkt

Digitalisierung und schulisches Lehren und Lernen

Joachim Griesbaum

Universität Hildesheim
Universitätsplatz 1
31141 Hildesheim
griesbau@uni-hildesheim.de

Zusammenfassung

Der Beitrag diskutiert Potentiale und Auswirkungen der Digitalisierung für schulisches Lehren und Lernen. Hierzu wird zunächst eine Übersicht über den Stand der Nutzung von Informations- und Kommunikationstechnologien in der Schule gegeben. Darauf aufsetzend werden Potentiale digitaler Technologien für Lernen und Unterricht anhand vorliegender empirischer Befunde diskutiert und Tendenzen des Bildungsmarktes geschildert. Ein Schwerpunkt stellt dabei der Einsatz mobiler Technologien dar.

Abstract

The article discusses the impact of digital technology for school-based teaching and learning. It starts with an overview of the current state of Information and Communication Technology (ICT) usage in schools in Germany. Following that, analyzing empirical findings, the potentials of ICT for school-based teaching and learning are discussed. In the discussion there is an emphasis on the usage of mobile technologies.

1. Einleitung

Diskutiert man die Potentiale der Digitalisierung für das Lernen, so kann man einerseits zwischen der Digitalisierung formaler Lernszenarien und dem nicht-formalen und informellen Lernen in digitalisierten Informationsumwelten differenzieren. Der erste Bereich fokussiert sich auf „gesteuertes“ Lernen in Lehre und Unterricht, der letztere lässt sich recht umfassend verstehen, bis hin zu den Auswirkungen der Digitalisierung für das allgemeine Informationsverhalten. Der folgende Beitrag konzentriert sich auf schulisches Lehren und Lernen. Auch in diesem engeren Themenfeld lassen sich neben dem Bereich des digitalen Unterrichts (a) auch Fragen des Informationsverhaltens in formalen Lernstrukturen (b) und des Informationsmanagements bei Lehrkräften und Schulen (c) behandeln. Nachfolgend werden zur Veranschaulichung die drei Bereiche kurz skizziert:

a) Digitaler Unterricht: behandelt Fragestellungen des Einsatzes von Technologien wie Whiteboards oder Tablets und Learning Apps im Unterricht und auch die damit verbundenen didaktischen Fragestellungen der (spielerischen, forschenden und/oder kollaborativen) Ausgestaltung von Lehr-/Lernszenarien.
b) Informationsverhalten in formalen Lernstrukturen: Fragen in diesem Bereich beinhalten z.B. die Nutzung von Facebook für die Erledigung von Hausaufgaben oder das Handyverbot im Unterricht.
c) Schulisches Informationsmanagement: Adressiert Fragestellungen zur Ressourcenallokation, Konfiguration und Organisation digitaler Technik in Schule und Unterricht. Dabei sind auch rechtliche und ergonomische Gesichtspunkte von Belang.

Der vorliegende Beitrag thematisiert zunächst die Verbreitung des digitalen „Klassenzimmers". Darauf aufsetzend werden die Wirkungspotentiale der Digitalisierung anhand grundsätzlicher theoretischer Erörterungen, der Darstellung spezifischer Fragestellungen zur individuellen Informationsverarbeitung und Ergonomie sowie der Schilderung spezifischer Fallstudien zum Einsatz digitaler Medien im Unterricht diskutiert. Anschließend werden Tendenzen des Bildungsmarktes angeschnitten und Aspekte des Informationsmanagements angeführt.

2. Verbreitung des digitalen „Klassenzimmers"

Für eine Bestandsaufnahme der digitalen Mediennutzung lassen sich die JIM- (Feierabend et al. 2014) und BITKOM-Studien (BITKOM, & LEARNTEC 2014) heranziehen. Hier wird eine Kluft zwischen privater und schulischer Medienausstattung und -nutzung deutlich. Während mehr als 50% der Schüler Computer für die Internetrecherche und Kommunikation zu Hausaufgaben heranziehen, lässt sich die Netznutzung in der Schule als eher punktuell bezeichnen. 12% der Schüler steht ein eigenes Notebook oder Tablet in der Schule zur Verfügung. Dezidierte Lernprogramme werden eher selten genutzt. Die IT-Ausstattung der Schulen wird mittlerweile von den Schülern mehrheitlich als schlecht bewertet. Der verstärkte Einsatz digitaler Medien wird gewünscht und als motivationsfördernd eingeschätzt. Die Lehrer werden nicht als technikfeindlich eingestuft, dennoch wird hier ein Weiterbildungsbedarf gesehen. Gemäß der Darstellung der Website www.tablet-in-der-schule.de setzt derzeit nur ein sehr geringer Teil der insgesamt rund 34.000 Schulen in der Bundesrepublik regelmäßig Tablet- Computer ein. Insgesamt zeigt sich also ein geringer Digitalisierungsgrad an deutschen Schulen und, zumindest auf Seiten der Schüler, ein gefühlter Nachholbedarf.

3. Wirkungspotentiale der Digitalisierung

Was sind nun die Wirkungspotentiale der Digitalisierung? Der Einsatz bzw. die Anwendung digitaler Medien im Unterricht lässt sich hinsichtlich der erzielten (positiven) Effekte nicht pauschalisieren. Insofern führen Fragestellungen oder Argumente, die Medieneinsatz unreflektiert mit dem Nicht-Einsatz von Medien kontrastieren – z.B. „Tabletunterricht" vs. „Non-Tablet-Unterricht" – in die Irre. Vielmehr sind die kontextuellen Bedingungsfaktoren der jeweiligen Lernumgebung zu berücksichtigten. Herzig (2014:9) argumentiert als Wirkungsfaktoren im Unterricht die Lernenden (z.B. Vorwissen, kognitive Ressourcen, Einstellungen, ...), Unterrichtsprozesse (Ziele, Inhalte, didaktische Struktur etc.), Lehrpersonen (Fachwissen, mediendidaktische Kompetenz usw.) und schließlich Digitale Medien (Ziele, Inhalte, Darstellungsform, Interaktivität u.a.m.). Dies bedeutet, es lassen sich positive Effekte spezifischer Konfigurationen argumentieren aber nicht pauschalieren. Im Folgenden sollen nun Wirkungsaspekte der Digitalisierung bzgl. Informationsverarbeitung und die Gestaltung von Unterricht diskutiert werden.

3.1 Informationsverarbeitung und Ergonomie

Ein grundlegender Mehrwert der Digitalisierung lässt sich in einer verbesserten individuellen Informationsverarbeitung argumentieren. Multimodalität und –codalität, d.h. die Kombination von auditiven und visuellen Medien oder Text und Bild, bei der Darstellung von Inhalten befördern deren Merkfähigkeit (Mayer 1997). Dabei sind die räumliche Nähe und Synchronizität der Mehrfachvariation entscheidend. Welche Effekte treten beim Lesen elektronischer Texte auf? Hier weisen vor allem ältere Studien auf eine geringere Lesegeschwindigkeit (Mayes et al. 2001) und reduzierte Genauigkeit (Dillon 1992) hin. Eine aktuelle Studie von Subrahmanyam et al. (2013), untersucht das Leseverhalten von 120 Studierenden auf Tablets, Laptops und Papier. Dabei werden verschiedene Bedingungen (Multitasking vs. Fokussiert und einfache vs. anspruchsvolle Textpassagen) hinsichtlich ihrer Auswirkungen auf Lesedauer und Textverständnis untersucht. Im Ergebnis zeigt sich kein Medieneffekt bzgl. Lesedauer und Textverständnis. Multitasking wirkt sich aber über alle Medien aus und bewirkt eine längere Lesedauer. Das Textverständnis wird auch hier nicht beeinträchtigt. Multitasking ist in den Computerbedingungen stärker ausgeprägt. Aus kognitiver Perspektive kann die Digitalisierung damit tendenziell Vorteile für die Informationsverarbeitung bewirken. In Bezug auf ergonomische Aspekte lassen sich sowohl Vor- als auch Nachteile argumentieren. Zunächst ist die Portabilität ein großer Vorteil. Notebooks, Tablets und auch Smartphones, gestatten die „Mitnahme" nahezu beliebiger „Wissensmengen". Folgende Abbildung

veranschaulicht zugleich mögliche Optionen, die Darstellung von Text anzupassen.

Abbildung 8: Textdarstellung auf mobilen Endgeräten.
Links Originalseite. Rechts drei alternative Darstellungen auf einem mobilen Endgerät.

Erstaunlich an der aktuellen Diskussion zum Einsatz mobiler Endgeräte ist, dass ergonomische und gesundheitsrelevante Aspekte bislang kaum thematisiert werden. Von der Erhöhung der „On-Screen-Time“ und möglichen Folgen für das Sehvermögen abgesehen, werden auch mittel- und langfristige physiologischen Folgen wie der sogenannten Tablet-Nacken (Young et al. 2012) aus der Diskussion bislang weitgehend ausgeblendet. Hinsichtlich ergonomischer Aspekte ist es deshalb anzuraten, nicht nur den konvenienten Informationszugriff zu argumentieren, sondern auch eher mittel- und langfristig zum Tragen kommende gesundheitliche Aspekte mit zu betrachten.

3.2 Fallstudien zum digitalen Unterricht

3.2.1 Knowledge-Building-Environments (Scardamalia & Bereiter 2003)

Erstmals 1983 eingesetzt und seit 1986 rund 20 Jahre im schulischen Betrieb verwendet, stellte CSILE/Knowledge Forum ein Projekt mit breitem Erfahrungshintergrund dar. Das Projekt verfolgte einen pädagogischen Ansatz der forschenden Lerngemeinschaft und implementierte so Ansätze des Computer Supported Collaborative Learning (CSCL) bereits seit den frühen 1980er Jahren. Ausgehend von Fragestellungen − z.B. „Ursachen von Umweltverschmutzung“ − formulieren Lernende ihre Ideen, Thesen, Fragen und führen alleine oder in Kleingruppen spezielle Aufgaben wie etwa eine Recherche nach einschlägigen Referenzen durch. Im Prozess des Knowledge Building erweiterte sich durch die Beiträge und die darauf folgenden Reaktionen schrittweise das gemeinsame Wissen. Insgesamt wies das Projekt früh auf die Potentiale und lernbezogenen Mehrwerte digitaler Technologien für das ge-

meinsame Erforschen von Wissen hin. Die folgende Abbildung zeigt einen Screenshot von Knowledge Forum.

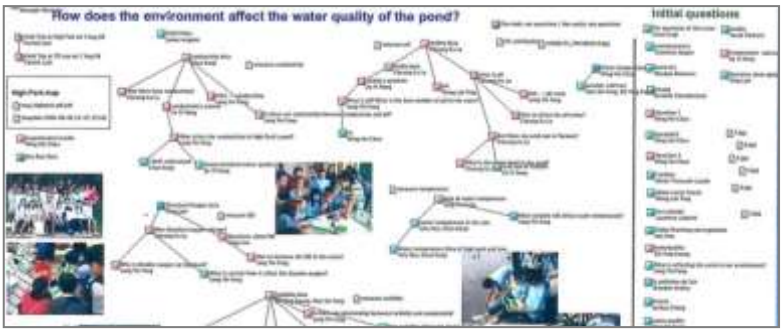
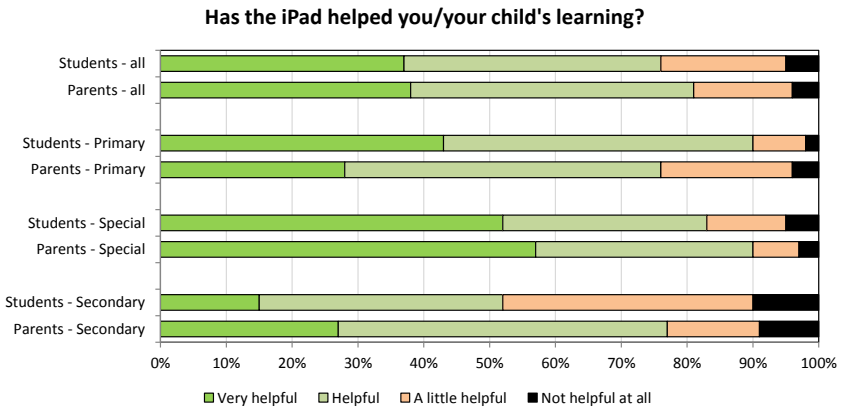


Abbildung 9: Diskussion in Knowledge Forum, Quelle: http://patricklamhc.wikispaces.com/mobile%20learning%202008

3.2.2 iPads for Learning (Department of Education and Early Childhood Development 2011)

Die gegenwärtige Diskussion zur Digitalisierung schulischen Lehrens und Lernens fokussiert sich stark auf die Einbindung mobiler Endgeräte, vor allem Tablets, in den Unterricht. Für eine erste Näherung ist es lohnenswert, einen Blick auf eine Initiative des Department of Education and Early Childhood in Australien zu werfen. Im Rahmen eines Pilotvorhabens wurden 2011 an zehn verschiedenen Schulen 666 iPads eingeführt. Die empfundenen Auswirkungen wurden bei den Anspruchsgruppen (Schüler, Lehrer, Eltern) nach einer einjährigen Testphase erfragt. Die Ergebnisse sind durchweg sehr positiv und deuten einen Wandel des Unterrichts in Richtung selbstgesteuertes Lernen und Kollaboration an. Folgende Abbildung zeigt eine der Ergebnisillustrationen in Bezug darauf, wie hilfreich das iPad für das Lernen eingeschätzt wird.



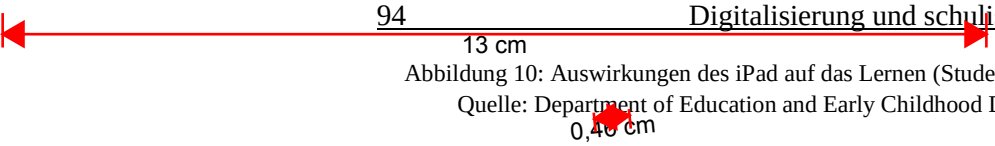


Abbildung 10: Auswirkungen des iPad auf das Lernen (Students: n=475, Parents: n=172, Quelle: Department of Education and Early Childhood Development 2011: 21

Erstaunlicherweise wird das iPad von „primary students“ (Primarstufe) häufiger als hilfreich eingestuft als von den älteren „secondary students“ (Sekundarstufe). Demnach lässt sich weniger ein Mindestalter für den Einsatz von Tablets argumentieren.

3.2.3 iPads in Early Childhood (Beschorner & Hutchison 2013)

Diesem Gedanken folgen auch Beschorner & Hutchison (2013). Sie untersuchen, wie der Einsatz von Tablets die Ausbildung von Schreib- und Lesefähigkeiten im Vorschulalter unterstützen kann. Für ihre Untersuchung werden zwei Kindergartenklassen mit Vier- und Fünfjährigen ausgewählt. Die Tabletnutzung der Kinder wird über 7 Wochen beobachtet. Im Ablauf der 7 Wochen wurden auf den Geräten immer wieder neue Apps installiert. Die Autoren kommen zu folgenden Ergebnissen: Die Kinder beschäftigen sich intensiv mit den Geräten, schauen gerne den anderen zu und geben Tipps. Sie sind kreativ und zeichnen Bilder, benutzen sogar die Tastatur. Insgesamt konstatieren Beschorner & Hutchison (2013) neben dem Erwerb erster Schreibfähigkeiten (z.B. Wörter nachmalen) einen vermehrten Austausch durch verstärkte Interaktion. Dies veranschaulicht auch das folgende Zitat: „*…both teachers suggested that the communication between children […] was the biggest difference […]. Mrs. Timmons noted that, even when working individually […], children would still engage in meaningful conversations with the children around them…*“ (ebd.: 22).

3.2.4 Mobiles Lernen in Hessen (Bremer & Tillmann 2014)

Bremer und Tillmann (2014) untersuchen begleitend ein Pilotprojekt zum Tableteinsatz an 6 Schulen in Hessen (2.-5. Klasse). Im Rahmen des Projekts werden Tablets primär als Gestaltungswerkzeuge genutzt, weniger als Medium zur Vermittlung von Inhalten. Vorliegende Befragungsergebnisse weisen auf eine hohe Akzeptanz auf Seiten der Schüler hin, die ergänzend ihren Lernerfolg bei der Arbeit mit dem iPad mehrheitlich als sehr hoch einschätzen.

3.2.5 Stanford Mobile Inquiry-based Learning Environment (SMILE) (Seol et al. 2011)

Im Rahmen des SMILE-Projekts untersuchen Seol et al. (2011) die Akzeptanz eines fragegeleiteten Lernszenarios, in dem Smartphones genutzt werden. Schüler erstellen eigene Fragen, zu den im Unterricht behandelten Themen, die dann von anderen Schülern beantwortet werden. Die erstellten Fragen werden durch die Mitschüler auch bewertet. Hier werden mobile Technologien also dazu genutzt um „zwischendurch“ eine konstruktivistische

und spielbasierte Komponente in den Unterricht einzubauen, die zugleich eine höhere Aufmerksamkeit bei der Darstellung des Themas bewirkt. Die Ergebnisse zeigen eine hohe Akzeptanz auf Seiten der Schüler. Diese erachten das Spiel vor allem für die Überprüfung von Lerninhalten als hilfreich.

3.2.6 Tablet-PC im Physikunterricht (Bresges et al 2013)

Bresges at al. (2013) vergleichen die Gruppenarbeit im Physikunterricht mit und ohne Tablet in einer 9. Klasse an einer Gesamtschule. Die Schüler führen ein Experiment durch, um die Schwerkraft zu erforschen. Die Versuchsgruppe soll ihre Arbeit mit dem Tablet dokumentieren und auswerten, die Kontrollgruppe mit Arbeitsblättern. Im Ergebnis zeigt sich eine verbesserte Gruppenkommunikation in der Versuchsgruppe. Bei der Tabletgruppe ist die Interaktion untereinander stärker und soziales Faulenzen geringer ausgeprägt. Gemäß den Ergebnissen eines Multiple-Choice-Tests nach dem Experiment ist der Lernzuwachs in der Tabletgruppe homogener, kein Schüler erreicht weniger als 50% aller Punkte. Insgesamt weist also auch diese Untersuchung darauf hin, dass der Einsatz digitaler Technologien sich bzgl. des Verhaltens im Lernprozess als auch des Ergebnisses sehr positiv auswirken kann.

3.2.7 Einsatz personalisierter iPads im Unterricht (Ludwig et al. 2011)

Ludwig et al. (2011) evaluieren ein dreimonatiges Pilotvorhaben zur Einführung von Tablets in den Klassen 12 und 13 mit insgesamt 16 „Tablet-Schülern". Dabei werden die Geräte von den ausgewählten Schülern selbstständig genutzt. Die Daten wurden durch Online-Tagebücher (in einem gemeinsamen Weblog), Befragungen und Unterrichtsbeobachtung erhoben. Im Kern sind auch hier die Befunde sehr positiv und weichen über eine engere Unterrichtsbetrachtung hinaus. So zeigen sich Vorteile bei der Informationsrecherche und des unmittelbarem Materialzugriffs. Auch werden die Geräte zur Unterstützung von Gruppenarbeit eingesetzt. Insgesamt werden die Geräte als „*Schaltzentrale des persönlichen Wissensmanagements*" eingestuft, zugleich verwischt aus Sicht der Schüler, die Grenze von schulischen und privaten Inhalten.

3.2.8 Zwischenfazit

Die Fallstudien veranschaulichen mögliche Vorteile der Digitalisierung in Bezug auf Lernpraktiken wie exploratives, kreatives und wissensgenerierendes Lernen und verstärkte Kooperation. Die Einführung mobiler Endgeräte scheint mit einem hohen motivationalen Effekt verbunden zu sein. Ob hier ein Neuigkeitseffekt vorliegt, bleibt zunächst offen. Generell lässt sich eine höhere „Mächtigkeit" auf Seiten der Lehrenden und Lernenden bei der (Aus-) Gestaltung und Durchführung von Lernszenarien konstatieren. Die Ubiquität des Informationszugriffs, die sozial unbeschränkten Kommunikati-

onsoptionen und die bei mobilen Endgeräten erweiterten Interaktionsoptionen schaffen vielfältige Potentiale eines effektiveren und effizienteren Informationsmanagements und erweitern die „Lernumwelt“ grundsätzlich. Betrachtet man die angeführten Studien etwas genauer so werden insbesondere lernbezogene Kernkompetenzen, wie Selbststeuerungsfähigkeit, Wissensgenerierung, Kommunikationskompetenz und in der Umsetzung dieser, eine Verschiebung hin zu schülerzentriertem Unterricht als positive Folge der Digitalisierung angedeutet.

Vor diesem Hintergrund ist die derzeitige Ausbildung eines Ökosystems sogenannter Lernapps spannend zu betrachten. Diese entsprechen in einer ersten Sicht oftmals in hohem Maße lerntheoretischen Anforderungen bzgl. der Realisierung spielbasierter und adaptiver Lernszenarien. Nachfolgende Abbildung zeigt ein solches Lernspiel, indem die Spieler die einsilbigen Wörter auswählen müssen.

Abbildung 11: „Emil und Pauline“ Lernspiel, vgl. http://www.emil-und-pauline.de/shop/neu/product/emil-und-pauline-auf-dem-hausboot-20-neu.html

Vordergründig werden durch solche Apps Optionen personalisierten Lernens befördert. Dennoch stellt sich die Frage, ob eine derartige *Appucation* erstrebenswert ist. So warnt Kohn (2014) vor altem Wein in neuen Schläuchen, in dem nach wie vor die Vermittlung von (Test-Score-relevanter) Information und nicht die Konstruktion von Bedeutung im Mittelpunkt steht. Laufen wir also bei aller Euphorie für die Vorteile der Digitalisierung auch gleichzeitig Gefahr alte Unterweisungspraktiken zu revitalisieren? Wissenschaftliche Evidenz lässt sich hier nicht anbringen. Es lässt sich ebenso eine Vielzahl von Apps und Webseiten finden, welche kreative, kommunikative und wissensgenerierende Lernmethoden unterstützen. Entscheidend ist also die Fra-

ge, welche Lerntechnologien wie genutzt werden. Die Rolle der Lehrkräfte bei der Gestaltung von Lernszenarien bleibt dabei nach wie vor zentral.

4. Tendenzen des Bildungsmarktes

Wie mit dem Begriff der *Appucation* schon angedeutet wandelt sich auch der Bildungsmarkt. Dies sei im Folgenden nur kurz und punktuell angerissen. Zunächst zeigt sich dies im Bereich der Anbieter von Bildungsmaterialien. Hier findet im kommerziellen Bereich ein Wandel vom klassischen „Schulbuchverlag“ hin zu Dienstleistern für Technologiemanagement und Digitale Lehrmaterialien statt. Es etablieren sich neue Lizenz- und Preismodelle. Beispiele stellen etwa die Online-Plattform Scook.de, die als digitaler Online-Aggregator wirkt oder die Firma Snappet (www.dasgrundschultablet.de) dar. Snappet stellt eine technische und inhaltliche „Komplettlösung für das Lernen mit Tablets an Grundschulen“ bereit. Neben diesen Entwicklungen auf dem kommerziellen Markt etablieren sich zunehmend auch im schulischen Bereich freie Bildungsmaterialien (Open Educational Resources: OER). So bietet die Zentrale für Unterrichtsmedien im Internet (Zum.de) eine umfangreiche Sammlung von OER für Unterricht und Lehrerausbildung. Auf internationaler Ebene strebt z.B. die Fuse School (about.me/TheVirtualSchool) an, eine Sammlung von mehreren tausend frei zugänglichen und kostenlosen Lernvideos aufzubauen.

Derzeit (Stand: 07.06.2015) finden sich auf dem entsprechenden Youtube-Kanal 410 Videos (http://youtube.com/user/virtualschooluk). Insgesamt zeichnen sich mit diesen Entwicklungstendenzen umfangreiche Änderungen bzgl. der technologischen und inhaltlichen Infrastruktur schulischen Unterrichts ab. Lehren und Lernen in der Schule wird so auch zu einen soziotechnischen Gestaltungsfeld, in dem Fragen der Bereitstellung und Gestaltung der Informationsinfrastruktur relevant werden. Die technische und soziale Ausgestaltung Schulischen Lehrens und Lernens bedarf daher auch in immer stärkerem Maße eines systematischen Informationsmanagements.

5. Aspekte des Informationsmanagements

Welches sind nun die wesentlichen Aspekte eines schulischen Informationsmanagements? Kernthemen der derzeitigen Diskussion stellen u.a. Medienkonzepte und Technologiemanagement für den Unterricht, die Gestaltung der Informationsinfrastruktur in der Schule, sowie rechtliche und sicherheitsbezogene Aspekte der Techniknutzung dar (Lehrer Online o.J.). Nachfolgend einige der Fragestellungen, die konkret zu adressieren sind: Was ist das Ziel

der Mediennutzung an der Schule? Was soll die schulische Infrastruktur leisten? Welche Einsatzszenarien gibt es für den Unterricht? Wer sind die Akteure und welche Handlungsnormen werden angelegt? Wie werden Ressourcen allokiert? Im Netz finden sich vielfältige Hilfestellungen Handreichungen und Informationsangebote zum Technologieeinsatz und Unterrichtsszenarien.[11] An dieser Stelle wird dafür plädiert, das Themenfeld konzeptuell auch aus der Perspektive eines systematischen Informationsmanagement (vgl. z.B. Krcmar 2010) zu fassen. Dies gestattet es, die vielfältigen anfallenden Einzelfragen systematisch anzugehen und zu strukturieren, ohne sich in den Einzelheiten zu verlieren. Des Weiteren ist es möglich, von vorliegenden Erkenntnissen zum Beispiel zur Technologieakzeptanz und des Change Management bei der Gestaltung und Einführung von Technologien und Systemen in der Schule zu profitieren.

6. Zusammenfassung und Ausblick

Welche Auswirkungen und Potentiale der Digitalisierung auf schulisches Lehren und Lernen lassen sich nun aus einer Gesamtperspektive zusammenführen? Zunächst wird deutlich, dass die Wirkungseffekte aus der Kombination vielfältiger kontextueller Faktoren resultieren. Technikzentrierte Sichtweisen greifen zu kurz. Vielmehr sind die jeweiligen Faktoren im Zusammenwirken von Lehrenden, Lernenden und Unterrichtsprozessen in die Betrachtung mit einzubeziehen. Des Weiteren wird in Bezug auf die individuelle Informationsverarbeitung und Ergonomie erkennbar, dass digitale Informationsverarbeitung Vorteile aufweisen kann und Lesen auf Bildschirm, dem Lesen auf Papier nicht a priori unterlegen ist. Zugleich wird gezeigt, dass langfristige physische Auswirkungen, die durch die spezifischen ergonomischen Eigenschaften bzw. durch die Handhabung von mobilen Geräten verursacht werden, derzeit kaum diskutiert werden. Weiterhin wird sichtbar, dass die Digitalisierung, insbesondere die Nutzung mobiler Endgeräte, sich aus didaktischer und lerntheoretischer Perspektive sehr positiv auswirken kann. Dabei ist nach wie vor die Kompetenz der Lehrkräfte entscheidend dafür, welche Effekte erzielt werden. Die behandelten Fallstudien zeigen zwar klar eine Tendenz zum schülerzentrierten Lernen. Die kurze Argumentation zu einer *Appucation* verdeutlicht aber auch, dass in Form von direktiv wissensvermittelnden Apps eine Art „entmenschlichte“ Unterweisung praktiziert werden kann. Schließlich zeigt sich am Wandel des Bildungsmarktes,

[11] U.a. auf den Bildungsservern, z.B. dem Hessischen Bildungsserver http://medien.bildung.hessen.de/service_medien/mke/index.html (letzter Zugriff 007.06.2015). Zu Unterrichtsszenarien z.B. Thissen (2013: 96f).

dass sich nicht nur der Unterricht ändert, sondern das gesamte Ökosystem Schule. Dabei werden insbesondere auch grundlegende Fragen der Gestaltung der Informationsinfrastruktur virulent. Die Digitalisierung führt also nicht nur zu einem unterrichtsbezogenen E-Learning, sondern ist deutlich weiter gefasst und betrifft das gesamte schulische Informationsmanagement.

7. Literaturverzeichnis

Beschorner, B., and Hutchison, A (2013). iPads as a Literacy Teaching Tool in Early Childhood. International Journal of Education in Mathematics, Science and Technology, 1, 1, 16-24.

Bremer, C.; Tillmann, A. (2014). Mobiles Lernen in Hessen: Erste Ergebnisse zum Einsatz von Tablets an hessischen Grundschulen. Vortrag VDE Workshop Digitale Medien in Lehre und Forschung. Online verfügbar unter http://www.medienhaus-frankfurt.de/MOLE/wp-content/uploads/2014/11/Vortrag_Mole_VDE_2014.pdf , letzter Zugriff 05.06.2015.

Bresges, A., Beckmannn, R., Schmoock, J., Quast, A., Schunke-Galley, J., Weber, J., Firmenrich, D.; Beckmann, R.; Kreiten, M. (2013). Das „Reichshofer Experimentierdesign "zur Entwicklung und Überprüfung des Einsatzes von iPad oder anderen Tablet-PC im Physikunterricht. PhyDid B-Didaktik der Physik-Beiträge zur DPG-Frühjahrstagung.

BITKOM, & LEARNTEC (Eds.) (2014, December 9): Digitale Schulen - vernetztes Lernen. Online verfügbar unter http://www.bitkom.org/files/documents/BTIKOM_Charts_PK_Digitale_Schule_09_12_2014.pdf, letzter Zugriff 05.06.2015.

Department of Education and Early Childhood Development (2011). iPads for Learning. In Their Hands Trial. Evaluation Report. Sunbury

Dillon, A. (1992). Reading from paper versus screens: A critical review of the empirical literature, in: Ergonomics, 35, 1297–1326.

Feierabend, S.; Plankenhorn, T. Rathgeb, T. (2014). JIM 2014. Jugend, Information, (Multi-)Media Basisstudie zum Medienumgang 12- bis 19-Jähriger in Deutschland, Medienpädagogischer Forschungsverbund Südwest (Hg), Stuttgart.

Herzig, B. (2014). „Wie wirksam sind digitale Medien im Unterricht?". Online verfügbar unter http://www.vielfalt-lernen.de/wp-

content/uploads/2014/09/DigitaleMedienUnterricht_final.pdf, letzter Zugriff 05.06.2015.

Kohn, A. (2014). Four Reasons to Worry About "Personalized Learning",. Online verfügbar unter http://www.alfiekohn.org/blogs/four-reasons-worry-personalized-learning/, letzter Zugriff 13.03.2015.

Krcmar, H. (2010). Informationsmanagement, Springer, Berlin.

Lehrer Online (o.J.). Dossier Medienkonzept. Online verfügbar unter http://www.lehrer-online.de/medienkonzept.php, letzter Zugriff 07.06.2015.

Ludwig, L., Mayrberger, K., & Weidmann, A. (2011). Einsatz personalisierter iPads im Unterricht aus Perspektive der Schülerinnen und Schüler. In Delfi Workshop 2011: 2. Workshop „Lerninfrastruktur in Schulen: 1:1-Computing". Fachtagung vom 05.–08. September 2011 an der TU Dresden, S.7-17.

Mayer, R.E. (1997). „Multimedia Learning Are We Asking the Right Questions?" Educational Psychologist (32) 1 1997. 1–19.

Mayes, D. K., Sims, V. K., & Koonce, J. M. (2001). Comprehension and workload differences for VDT and. International Journal of Industrial Ergonomics, 28, 367–378.

Scardamalia, M.; Bereiter, C. (2003). Knowledge-Building-Environments: Extending the limits of the possible in education and knowledge work. In: Encyclopedia of distributed learning. DiSte-fano, A.; Rudestam, K. E.; Silverman, R. (eds.). Thousand Oaks, CA: Sage Publications.

Seol, S.; Sharp, A.; Kim, P. (2011). Stanford Mobile Inquiry-based Learning Environment(SMILE): using mobile phones to promote student inquires in the elementary classroom, in: Proceedings of WORLDCOMP'11: The 2011 World Congress in Computer Science, Computer Engineering, and Applied Computing. Online verfügbar unter https://gse-it.stanford.edu/sites/default/files/worldcomp11_SMILE.pdf, letzter Zugriff 06.06.2015.

Subrahmanyam, K.; Michikyan, M.; Clemmons, C.; Carrillo, R.; Uhls, Y. T.; Greenfield (2013). Learning from Paper, Learning from Screens: Impact of Screen Reading and Multitasking Conditions on Reading and Writing among College Students. International Journal of Cyber Behavior, Psychology and Learning, 3(4), 1-27, October-December 2013.

Thissen, F. (2013). Mobiles Lernen in der Schule. Online verfügbar unter http://www.frank-thissen.de/ibook_gut.pdf, letzter Zugriff 07.06.2015.

Young, J.G.; Trudeau, M.; Odell, D.; Marinelli, K.; Denerlein, J.T. (2012). Touch-screen tablet user configurations and case-supported tilt affect head and neck flexion angles, in: Work, 41, 81-91.

Ein Lernprogramm für Zulu-Possessivkonstruktionen

Erweiterung eines bilingualen monodirektionalen elektronischen Wörterbuchs

Alexandria-Barbara Sanasi

Universität Hildesheim
Universitätsplatz 1
31141 Hildesheim
sanasi@uni-hildesheim.de

Zusammenfassung

Zulu zählt zu den afrikanischen Bantusprachen, welche ein komplexes Flexionssystem aufweisen. Das Anwenden der z.T. sehr komplexen grammatischen Regeln erschwert das Erlernen einer solchen Sprache, weshalb ein Bedarf an zusätzlichen Online-Hilfsmitteln zum Üben spezifischer Konstruktionen besteht. Das hier vorgestellte Lernprogramm für Possessivkonstruktionen baut auf einem elektronischen Wörterbuch auf und integriert verschiedene didaktische Feedbackstrategien aus dem Präsenzunterricht. Ebenso werden zusätzliche Hilfen angeboten, auf die der Lernende bei Bedarf zurückgreifen kann. Das Lernprogramm verbindet lexikografisches Wissen (aus dem elektronischen Wörterbuch) mit didaktischen Elementen des Lehrens und Lernens und ist auf mehreren Ebenen nutzerorientiert gestaltet.

Abstract

Zulu belongs to the Bantu languages, which are known for their rich agglutinating morphological structure. The use of partially complex grammar rules makes it difficult to learn such languages. This is why, the demand for additional online tools for training specific constructions, exists. The here proposed learning program for possessive constructions is based on an electronic dictionary and integrates didactic feedback strategies, known from **F**oreign **L**anguage **A**cquisition and **T**eaching (**FLA/T**). Furthermore, the learning program provides additional help functions, which can be accessed by the user on demand. The learning program links lexicographic knowledge with didactic elements from FLA/T and follows an user-oriented design.

1. Einleitung

Zulu ist, wie Japanisch oder Finnisch, eine agglutinierende Sprache und zählt zu den Nguni-Sprachen, welche ihrerseits zu der süd-westlichen Bantuzone gezählt werden (vgl. Bosch/Pretorius/Fleisch, 2008, 70). Diese Sprachen weisen ein komplexes Flexionssystem auf und kennzeichnen sich durch ein nominales Klassensystem mit Kongruenz-Elementen. Aufgrund ihrer komplexen Grammatik erweisen sich die Bantusprachen als interessant und herausfordernd im Sinne der Gestaltung von Lernprogrammen, die Lernende unterstützen sollen.

In diesem Beitrag stellen wir den aktuellen technischen Stand eines Lernprogramms für Zulu-Possessivkonstruktionen vor, welches im Rahmen des *SeLA*-Projektes[12] (*Scientific eLexicography for Africa*) in Kooperation mit Zulu-Lehrenden an der südafrikanischen Fernuniversität *University of South Africa*, UNISA, erstellt wurde.

Das Lernprogramm übernimmt verschiedene Feedbackmethoden aus dem Präsenzunterricht und bietet die Möglichkeit, zusätzliche Hilfen, wie zum Beispiel grammatische Tabellen, Links oder andere externe Quellen, aufzurufen (*information on demand*).

Das Lernprogramm basiert auf einem elektronischen Wörterbuch, dem *eZulu Dictionary* (vgl. Bosch/Faaß, 2014) und verlinkt dadurch lexikografisches Wissen mit didaktischen Elementen des Lehrens und Lernens. Weiterhin ist es auf mehreren Ebenen nutzerorientiert gestaltet.

Der Beitrag gliedert sich wie folgt: Zunächst werden der Rahmen der Forschungen und die Motivation für die vorliegenden Arbeiten dargestellt (Abschnitt 2). Abschnitt 3 beschäftigt sich mit dem didaktischen Element der Feedbackstrategien aus dem Präsenzunterricht, während im darauffolgenden Abschnitt 4 die Umsetzung in aktuellen Online-Medienangeboten diskutiert wird. In Abschnitt 5 thematisieren wir die Bildung von Zulu-Possessivkonstruktionen und geben einen kurzen Einblick in die Funktionsweise des *eZulu Dictionary*. Abschnitt 6 beschäftigt sich mit der technischen Umsetzung des Lernprogramms für Zulu-Possessivkonstruktionen und die Integrierung der didaktischen Feedbackstrategien. Schließlich werden in Abschnitt 7 die Ergebnisse der vorangegangenen Abschnitte zusammengefasst und ein kurzer Ausblick über bevorstehende Arbeiten gegeben.

---

[12] Näheres unter: http://www.uni-hildesheim.de/iwist-cl/projects/sela/. Letzter Zugriff: 15.6.2015

2. Motivation und Zielsetzung

Die südafrikanische Fernuniversität University of South Africa (UNISA) lehrt Zulu als Fremdsprache. Die dortigen Fernstudierenden sind bisher hauptsächlich auf gedrucktes Lehrmaterial, die so genannten *Study Guides* (Studienbriefe), angewiesen. Jedoch steigt der Bedarf nach zusätzlichen Online-Werkzeugen zum Vertiefen des Gelernten und zum praktischen Üben komplexer grammatischer Strukturen.

Das hier vorgestellte Lernprogramm für Zulu-Possessivkonstruktionen, im Folgenden bezeichnet als *Grammar Trainer*, versucht, diesen Bedarf zu decken. Der *Grammar Trainer* baut auf dem bereits existierenden *eZulu Dictionary* auf (vgl. Bosch/Faaß, 2014) und stellt eine Implementierung bekannter didaktischer Elemente (z.B. Feedbackstrategien) aus dem Präsenzunterricht dar.

3. Didaktische Methoden aus dem traditionellen (Fremdsprachen-)Unterricht

3.1 Lehr-/Lernsituation

Traditionelle Lehr-/Lernmethoden basieren auf „face-to-face"-Situationen (z.B im Klassenzimmer, Präsenzunterricht). Ein Lehrender vermittelt dem Lernenden ausgewählte Inhalte und gibt entsprechend der Lernerantwort positives oder korrigierendes Feedback. Die Methodik, wie Wissen vermittelt wird, kann hierbei variieren, ebenso die Art des gegebenen Feedbacks. Der Lehrende (in seiner Rolle als Experte) wählt sowohl bewusst das Lehrmaterial als auch die Art und Weise den Lernenden zu „loben" oder zu korrigieren. Das Angebot an traditionellen Lehr-/Lernmethoden ist groß; für die folgenden Ausführungen werden wir uns nur auf die verschiedenen Feedbackstrategien konzentrieren.

3.2 Feedbackstrategien

Der Begriff *Feedback* wird in verschiedenen technischen und wissenschaftlichen Domänen benutzt (vgl. Narciss, 2008, 125). In diesem Beitrag verstehen wir Feedback als Rückmeldung auf Lernerantworten und -handlungen, sowohl positiv als auch korrigierend. Dieses Verständnis korreliert mit der vorgeschlagenen Definition von Narciss (ebd., 127): *„Feedback is all post-response information that is provided to a learner to inform the learner on his or her actual state of learning or performance."* Die im folgenden vorge-

stellten Feedbackstrategien basieren auf Ellis (1997, vgl. Abschnitt 3.2.1) und Ferreira/Moore/Mellish, (2007, 392f., vgl. Abschnitt 3.2.2).

3.2.1 Positives Feedback

Positives Feedback kennzeichnet sich dadurch, dass den Lernenden Rückmeldung über die Fehlerfreiheit ihrer Antwort gegeben wird. Dies kann auf zwei Arten erfolgen („*Repeat the correct answer*“ oder „*Rephrase or Reformulate the correct answer*“), siehe Tabelle 1.

Strategie	***Erläuterung***	***Beispiel***
Repeat the correct answer	Der Lehrende wiederholt die Lernerantwort.	S(chüler):„Voortrekker Monument.“ L(ehrer): „Voortrekker Monument.”
Rephrase or Reformulate the correct answer	Der Lehrende wiederholt die Lernerantwort und fügt zusätzliche Informationen hinzu	S: „Berlin. “ L: “Berlin, Cologne and Düsseldorf are German cities with a high cultural diversity.”

Tabelle 1: Positive Feedbackstrategien

3.2.2 Korrigierendes Feedback

Korrigierendes Feedback kennzeichnet sich dadurch, dass den Lernenden Rückmeldung über die Fehler(art) der Antwort gegeben wird. Es kann zwischen zwei Typen unterschieden werden (vgl. Ferreria/Moore/Mellish, 2007, 392).

(a) **G**ive-**A**nswer **S**trategie (**GAS**) und
(b) **P**rompting-**A**nswer **S**trategie (**PAS**).

GASn kennzeichnen sich durch die direkte Vermittlung der korrekten Antwort, während sich PASn darauf fokussieren, die Lernenden zur korrekten Antwort hinzuleiten.

Give-Answer Strategie, GAS

Strategie	***Erläuterung***	***Beispiel***
Repetition of an incorrect answer	Der Lehrende wiederholt die falsche Antwort.	S: „The word ‚went' is in present tense." L: „Present tense?"
Recast of an incorrect answer	Der Lehrende formuliert die falsche Antwort in einem neuen Satz um, der die korrekte Antwort enthält.	S: „Angela Merkel is Germany's male chancellor." L: „That's right. Angela Merkel is the first female chancellor in Germany. "
Rephrasing of the wrong parts of a partially correct answer	Der Lehrende gibt die richtige Antwort, ohne den Lernerfehler zu wiederholen.	S: „The girl is went to school. " L: "*went* (?)"
Provision/ Completion of an correct answer	Der Lehrende gibt die richtige Antwort, wenn der Lernende diese nicht weiß.	S: „Yesterday we bought many things in the supermarket: eggs, flour, and…" L: "flour and sugar."

Tabelle 2: Korrigierende Feedbackstrategien - GAS

Prompting-Answer Strategie, PAS

Strategie	***Erläuterung***	***Beispiel***
Meta-linguistic cues	Der Lehrende gibt Hinweise, die auf die richtige Antwort deuten.	S: "The girl run around the school building." L: "Here you have to make use of the third person of 'run' in the present tense."
Request to clarify an incorrect answer	Der Lehrende stellt Fragen und impliziert damit, dass die Lernerantwort nicht korrekt ist.	S: „I would like to drink mullad wine. " L: "What? Do you really mean 'mullad'-wine?"
Elicit the correct answer	Der Lehrende ermutigt die Lernenden einen Satz oder eine Aufgabe zu vollenden. Es werden wie in *Metalinguistic cues* Hinweise gegeben.	L: „Warum freuen sich Kinder in Deutschland auf den 6.Dezember? " S: Because they believe that Saint Nicholas comes and brings presents." L: "And how do you say that in German?"

Tabelle 3: Korrigierende Feedbackstragien – PAS

4. Feedbackgestaltung

Das traditionelle Lehren und Lernen wird immer häufiger durch den Einsatz neuer Medien ergänzt bzw. angereichert. Dieses Prinzip wird als *Blended Learning* bezeichnet (vgl. Mandl/Kopp, 2006, 6) und integriert die Potentiale von Online-Medienangeboten mit traditionellen Lehrmitteln. Das zu übermittelnde Wissen nimmt stetig zu, was dazu führt, dass sowohl Lehrende als auch Lernende dazu gezwungen sind, auf neuere, flexiblere Methoden, wie zum Beispiel *E-Learning* (vgl. Condruz-Bacescu, 2014, 159fff.), umzusteigen.

Sanasi (2014) hat einen deskriptiven Vergleich von bekannten kostenlosen **O**nline **S**prach**L**ern**S**ystemen (**OSLS**) durchgeführt und überprüft, wie und ob Feedbackstrategien (siehe Abschnitt 3.1) aus dem Präsenzunterricht übernommen wurden. In diesem Vergleich konnten nur sechs verschiedene OSLS näher untersucht werden, daher kann eine Beurteilung nur im Rahmen der betrachteten Systeme stattfinden und nicht verallgemeinert werden. Eine detaillierte Übersicht zur Anwendung der oben angeführten Feedbackstrategien in den einzelnen Systemen wird in Tabelle 4 gegeben. Die nun folgende Legende dient zum besseren Verständnis von Tabelle 4.

- **Farbliche Hinterlegung**: orange = Positives Feedback, blau = Korrigierendes Feedback, GAS; grün = Korrigierendes Feedback, PAS
- **Feedbackstrategien:** [a] Repetition of the correct answer (RepCA), [b] Reformulate or Rephrase the correct answer (ReformCA), [c] Repetition of an incorrect answer (RepINCA), [d] Recast of the wrong parts of a partially correct answer (RecINCA), [e] Provision/Completion of a correct answer (ProvCA), [f] Meta-linguistic cues (MetaLing), [g] Request to clarify an incorrect answer (ReqINCA), [h] Elicitation of correct answer (ElicitCA)

	Babbel[13]	Busuu[14]	Duolingo[15]	Live Mocha[16]	phase-6[17]	Rosetta Stone[18]
RepCA	Ja	ja	ja	nein	nein	nein
ReformCA	Nein	nein	nein	nein	nein	nein
RepINCA	Ja	ja	ja	nein	nein	nein
RecINCA	Nein	nein	nein	nein	nein	nein
ProvCA	Ja	ja	ja	ja	ja	ja
MetaLing	Nein	nein	nein	nein	nein	nein
ReqINCA	Nein	nein	nein	nein	nein	nein
ElicitCA	Nein	nein	nein	nein	nein	nein

Tabelle 4: Übersicht angewandter Feedbackstrategien in einer Auswahl von online-Sprachlernsystemen

Weiterhin wurden andere Feedbackmethoden, die nicht aus dem Präsenzunterricht stammen, identifiziert. Diese Methoden arbeiten eher mit Visualisierung (zum Beispiel: farblichen Hervorhebungen, Gebrauch von speziellen Symbolen, etc.) als mit der Verbalisierung von Feedback. Ein Element aus dem Präsenzunterricht, welches den Online-Medienangeboten verloren geht, ist die Intonation und der Gebrauch von Mimik und Gestik, um Feedback zu vermitteln. Stattdessen werden Farben, Symbole oder auch Bilder dazu verwendet, den Lernenden auf die Fehlerfreiheit oder Richtigkeit der gegebenen Antwort hinzuweisen. Diese Methoden werden in Tabelle 5 detailliert aufgelistet.

[13] ***URL**: http://de.babbel.com/*
[14] ***URL**: https://www.busuu.com/de/*
[15] ***URL**: https://www.duolingo.com/*
[16] ***URL**: http://livemocha.com/*
[17] ***URL**: http://www.phase-6.de/*
[18] ***URL**: http://www.rosettastone.de/*

	Babbel	Busuu	Duolingo	Live Mocha	phase-6	Rosetta Stone
Hervorhebung (z.B. Unterlegung in einer anderen Farbe)	ja	Ja	Ja	ja	nein	Ja
Farbliche Markierung (z.B grün für korrekte, rot für falsche Antworten)	ja	ja	Ja	ja	nein	Ja
Verbales Feedback (z.B. „Das war richtig, gut gemacht!“ oder „Das war leider falsch…“)	nein	ja	Ja	nein	ja	nein
Symbole (z.B. Häkchen, Kreuz, etc.)	nein	ja	Ja	nein	nein	ja
Audio (z.B Signaltöne bei korrekter Auswahl)	ja	ja	Ja	ja	ja	ja

Tabelle 5: Visuelle Gestaltung von Feedback in der Selektion von online Sprachlernsystemen

Eine Zusammenfassung aller Methoden (Feedbackstrategien, visuelle Elemente, etc.), die beim Vergleich der sechs OSLS identifiziert wurden, ist in Sanasi (2014, 23f.) zu finden.

5. Lernprogramme für Zulu

Das *eZulu Dictionary* scheint bisher das einzige Online-Medium zu sein, welches Zulu-Lernende beim Sprachenerwerb unterstützt, speziell bei der Bildung von Possessivkonstruktionen.

Zulu zeichnet sich durch ein Nominalklassensystem aus. Jede Klasse besitzt neben anderen sprachlichen Elementen ein ihm zugeordnetes Kongruenz-Element. Zulu-Possessivkonstruktionen bestehen aus einer Abfolge von Morphemen der folgenden Art: „Besitztum Kongruenz-Element+Besitzer“: Das Kongruenz-Element und der Possessor verschmelzen dann zu einem orthografischen Wort. Für die Bildung von Possessivkonstruktionen gelten morpho-phonologische Regeln, die von den Nominalklassen der Possession und des Possessors bestimmt werden.

Das *eZulu Dictionary* übersetzt auf Basis lexikografischer Daten und morpho-phonologischer Regeln englische Possessivkonstruktionen nach Zulu. Die Übersetzung erfolgt hierbei schrittweise und wird dem Benutzer auf

Wunsch dargestellt. Der *Grammar Trainer* baut auf dem *eZulu Dictionary* auf, das seinerseits aus drei Modulen besteht: (a) einer Datenbank, welche die lexikografischen Daten (Übersetzungsäquivalente, nominale Klassen, etc.) beinhaltet, (b) morpho-phonologischen Regeln (repräsentiert in php-Skripten) und (c) einer grafischen Benutzerschnittstelle (GUI) bestehend aus html-Dateien.

Abbildung 1 zeigt die Ergebnisseite einer Übersetzung im *eZulu Dictionary*. Diese Ansicht erscheint, wenn der Benutzer die Option gewählt hat, sich die Regeln der Possessivbildung anzeigen zu lassen. Wählt der Benutzer diese Option nicht, erscheint nur die hell-blau hervorgehobene obere Zeile.

Abbildung 12: Schrittweise Übersetzung im *eZulu Dictionary* (Englisch nach Zulu)

6. Umsetzung

6.1 Prinzipien

Wie bereits in Abschnitt 5 beschrieben, baut der *Grammar Trainer* auf den lexikografischen Daten und den morpho-phonologischen Regeln des *eZulu Dictionary* auf (vgl. Abbildung 2). Die Datenbank wird um drei Tabellenein-

träge erweitert, um eine nutzerorientierte Gestaltung zu gewährleisten und um semantisch inakzeptable Possessivkonstruktionen[19] als Stimuli zu vermeiden. Die komplette Implementierung des *Grammar Trainers* ist dokumentiert in Sanasi (2015, 16-35). Hinzu kommen die Kategorien: (a) Level, zur Einstufung der Benutzerkenntnisse und (b) Möglichkeit Possession und/oder Possessor zu sein. In (a) sollen nur jene Wörter vom Trainer ausgewählt werden, die er Lernstufe des Lernenden entsprechen (z.B. Level 1 für Anfänger, Level 2 für Experten). Die Kategorie (b) dient dazu, die lexikografischen Daten nach potentiellen Possessionen und Possessoren zu filtern[20].

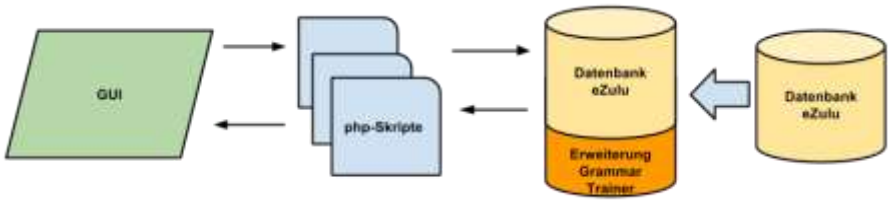


Abbildung 13: Struktureller Aufbau des *Grammar Trainers*

6.2 Feedbackstrategien

6.2.1 6.2.1 Positives Feedback

Der *Grammar Trainer* integriert sowohl (a) positives als auch (b) korrigierendes Feedback. Für (a) setzt er die Strategie [a] *Repeat the correct answer* und [b] *Rephrase or Reformulate the correct answer*, um. Ein Beispiel hierfür ist in Abbildung 3 zu sehen. Für die Beschreibung der Umsetzung der Feedbackstrategien werden wir uns in den folgenden Abbildungen auf die Übersetzung von *dog of man*, welche eine reguläre Bildung darstellt, stützen.

Abbildung 14: Positives Feedback für eine korrekte Übersetzung der Possession im *Grammar Trainer*

[19] Anmerkung: Possessivbildungen wie *letter of air* (*air's letter*) sollen nicht möglich sein.

[20] Anmerkung: Die Wahrscheinlichkeit, dass das Nomen „Essen" (Zulu: ukudla) eher Possession als Possessor ist, liegt höher als bei dem Nomen „Mann" (Zulu: indoda).

Im Fall, dass Possession (und Possessor) korrekt übersetzt werden, gibt der Trainer die positive Meldung aus: „*Well done! The possession of 'inja'. It belongs to noun class 09*". Dem Lernenden wird die Information der nominalen Klasse ausgegeben, welche in bestimmten Sonderfällen benötigt wird (mehr dazu in Bosch/Faaß, 2014, 740f.). In den übrigen Schritten der Übersetzung (vgl. hierzu Abbildung 1) erfolgt eine positive Rückmeldung nach Strategie [a] (vgl. Abbildung 4 und 5).

Abbildung 15: Positives Feedback für eine korrekte Identifizierung des Kongruenz-Elements im *Grammar Trainer*

Abbildung 16: Positives Feedback für die korrekte Vereinigung des Kongruenz-Elements mit dem Possessor im *Grammar Trainer*

Nach jedem erfolgreich abgeschlossenem Schritt werden die Antworten visuell festgehalten. Somit ergibt sich nach einer vollständig abgeschlossenen Übersetzung der englischen Konstruktion eine tabellenartige Zusammenfassung der übersetzten Elemente (siehe Abbildung 6).

1. Step: Translate **dog** (Possession)	dog is **inja**	class 09
2. Step: Translate **man** (Possessor)	man is **indoda**	class 09
3. Step: Determine the Possessive Concord (of **dog**)	Connecting element:	"ya"
4. Step: Link "**ya**" and **indoda**	"ya" + "indoda"	"yendoda"
Full translation	dog of man is	inja yendoda

Abbildung 17: Zusammenfassung aller sprachlichen Elemente nach einer korrekten Übersetzung im *Grammar Trainer*

6.2.2 Korrigierendes Feedback

Im Fall einer falsch gegebenen Antwort bietet der Trainer sowohl GA- als auch PA-Strategien an. Von den GAS setzt es die Strategie [c] und von PAS die Strategie [f] ein.

Wird die Possession oder der Possessor falsch übersetzt, so gibt es für den Trainer zwei Möglichkeiten zu reagieren:

(a) Ist die Lernerantwort partiell falsch, d.h. es befindet sich u.U ein Schreibfehler in der Antwort[21], prüft eine Levensthein-Funktion in dem entsprechenden php-Skript, welches Nomen dem eingegebenen Wort am ähnlichsten ist, und gibt es dem Lerner aus (vgl. Abbildung 7).
(b) Gibt der Lerner jedoch ein Wort ein, welches auch in der Datenbank existiert, aber nicht das gesuchte Nomen ist, so verweist der Trainer auf das *eZulu Dictionary* und bietet somit dem Lernenden die Möglichkeit, das Nomen nachzuschlagen.

Abbildung 18: Korrigierendes Feedback für ein eingegebenes Wort mit Tippfehlern im *Grammar Trainer*

Weitere korrigierende Feedbackmeldungen beziehen sich auf die inkorrekte Auswahl des Kongruenz-Elements oder auf die Nichteinhaltung der morphophonologischen Regeln. Der Benutzer kann darüber hinaus jederzeit externe Quellen wie z.B *isiZulu.net* aufrufen, um zusätzliche Hilfen zu suchen.

7. Resultat und Ausblick

Der *Grammar Trainer* für Zulu-Possessivkonstruktionen baut auf einfachen Feedbackstrategien aus dem Präsenzunterricht auf (vgl. Abschnitt 3.2). Für eine positive Rückmeldung nutzt der Trainer die Methode der (a) Wiederholung und (b) Anreicherung der Antwort mit zusätzlichem Wissen (*Repeat* und *Rephrase the correct answer*). Für korrigierendes Feedback bedient sich der Trainer sowohl GAS als auch PAS.

Zudem werden visuelle Methoden, die Sanasi (2014) im Vergleich verschiedener OSLS identifiziert hat, verwendet: (i) Die farbliche Hervorhebung von korrekten und inkorrekten Antworten (grün=korrekt, rot=inkorrekt) und (ii) verbales Feedback im Sinne einer „lobenden“ oder „korrigierenden“ Meldung (zum Beispiel: „*Well done!*“ oder „*Oh no, something went wrong…*“). Die Verwendung der Farben grün (positives Feedback) und rot (korrigierendes Feedback) sind mit der UNISA in Übereinkunft getroffen worden.

Weitere Arbeiten am *Grammar Trainer* betreffen (1) den Ausbau der bereits implementierten Feedbackstrategien (vgl. [f]) und weiterer visueller und

[21] Beispiel: Der Lerner will *inja* („Hund“) schreiben, aber er vertippt sich und schreibt *inya*.

auditiver Mittel (z.B Signaltöne, der Einsatz von bestimmten Symbolen) und (2) die Evaluierung des Trainers an der *University of South Africa*, UNISA, zum Zwecke der Optimierung und Implementierung weiterer didaktischer Mittel. Ebenso sollen die Datenbankeinträge des *Grammar Trainers* erweitert werden, sodass verschiedene Possessivkonstruktionen geübt werden können.

8. Literaturverzeichnis

Bosch, Sonja; Faaß, Gertrud (2014): Towards an Integrated E-Dictionary Application - The Case of an English to Zulu Dictionary of Possessives. In: Proceedings of the XVI EURALEX International Congress: The User in Focus: 739-747, Bolzano, July 15-19 2014. 739-747.

Bosch, Sonja; Pretorius, Laurette; Fleisch, Axel (2008): Experimental Bootstrapping of Morphological Analyser for Nguni languages. In: Nordic Journal of African Studies 17 (2), 66-88.

Condruz-Bacescu, Monica (2014): E-Learning/M-Learning – The new Trend in Foreign language Teaching. In: Professional Communication and Translation Studies 7(1-2). 159-166.

Ferreira, Anita; Moore, Johanna; Mellish, Chris (2007). A Study of Feedback Strategies in Foreign Language Classrooms and Tutorials with Implications for Intelligent Computer-Assisted Language Learning Systems. In: International Journal of Articial Intelligence in Education 17 (4), 389-422.

Mandl, Heinz; Kopp, Birgitta (2006): Blended Learning: Forschungsfragen und Perspektiven (Forschungsbericht Nr. 182). München: Ludwig-Maximilians-Universität, Department Psychologie, Institut für Pädagogische Psychologie.

Narciss, Susanne (2008): Feedback strategies for interactive learning tasks. In: Spector, J.M.; Merrill, M.D.; van Merrienboer, J.J.G.; Driscoll, M.P. (2008): Handbook of Research on Educational Communications and Technology Mahaw, NJ: Lawrence Erlbaum Associates. 125-144.

Sanasi, Alexandria-Barbara (2014): ICALL systems for Zulu. State-of-the art: An overview. Universtiät Hildesheim. December, ms., [BA-Projekt]

Sanasi, Alexandria-Barbara (2015): An ICALL application for Zulu: implementation of Zulu possessive construction. Universität Hildesheim, Februar, ms. [BA-Arbeit]

Wahrnehmung und Effektivität suchbezogener Werbung auf Smartphones

Alexa Domachowski, Joachim Griesbaum**, Ben Heuwing****

*Heise Media Service
Karl-Wiechert-Allee 10
30625 Hannover
Alexa.Domachowski@heise.de

**Universität Hildesheim
Universitätsplatz 1
31141 Hildesheim
griesbau@uni-hildesheim.de

***Universität Hildesheim
Universitätsplatz 1
31141 Hildesheim
heuwing@uni-hildesheim.de

Zusammenfassung

Der Beitrag schildert die Ergebnisse einer Untersuchung zur Wahrnehmung und Effektivität suchbezogener Werbung auf Smartphones. Hierzu absolvierten 20 Teilnehmer vier verschiedene Suchaufgaben. Eye-Tracking ermöglichte die Erfassung der Aufmerksamkeit für die Werbeanzeigen. Die Effektivität wurde durch das Selektionsverhalten (Klicks) gemessen und die individuelle Einschätzung der suchbezogenen Werbeanzeigen durch die Teilnehmer erfragt. Insgesamt weisen die Ergebnisse darauf hin, dass Nutzer auch auf Smartphones suchbezogene Werbung vermeiden.

Abstract

This paper examines the perception of Paid Listings within Google search results on smartphones. For that purpose, an exploratory study including 20 subjects that carried out four different search tasks was conducted. User perception of search results was measured through eye tracking and click tracking on the corresponding search engine result page. In addition, users were interviewed with regard to their subjective views on Paid Listings. Results indicate that similar to desktop search, users tend to avoid Paid Listings.

1. Suchbezogene Werbung im Web

Suchbezogene Werbeeinblendungen bilden den Kern des Geschäftsmodells der großen Universalsuchdienste. Der Marktführer Google generierte im ersten Quartal 2015 rund 17 Mrd. Dollar Umsatz (Google 2015). Alleine auf Google Sites (AdWords) wurden knapp 12 Mrd. Dollar, das sind 70% des

gesamten Umsatzes, verdient. Der Suchwortvermarktungsdienst AdWords bildet aus geschäftlicher Perspektive die finanzielle Grundlage der Websuche von Google, welche aus Nutzersicht kostenfrei ist. Die Effektivität der Suchwortanzeigen ist damit eine der zentralen Bedingungsfaktoren des gegenwärtigen Internets. Sie ist wiederum abhängig von der Wahrnehmung und Akzeptanz der Werbeeinblendungen durch die Nutzer der Suchmaschine. Dadurch, dass Nutzer die Information explizit anfragen und die Werbung darauf abgestimmt wird, ist davon auszugehen, dass die Nutzerreaktanz deutlich geringer ausfällt im Vergleich zu Display Advertising, welches Nutzer dazu bewegen soll, ihre aktuelle Tätigkeit zu unterbrechen. Studien weisen dennoch darauf hin, dass Nutzer die regulären organischen, auf Relevanzkriterien beruhenden Einträge auf den Ergebnisseiten der Suchmaschinen präferieren (Buscher et al. 2010; Jansen & Resnick 2006).

Mit der zunehmenden Diffusion des mobilen Internets ändern sich die Rahmenbedingungen für suchbezogene Werbung. Google selbst bestätigt den zunehmenden Trend zur mobilen Suche und schreibt im Mai 2015, dass in 10 Ländern (darunter die USA) von mobilen Geräten aus mehr Suchanfragen als von stationären Geräten durchgeführt werden (Dischler 2015). Der mobile Werbemarkt wird demnach zu einem primären Distributionskanal suchbezogener Werbung und damit zu einem zentralen Erfolgsfaktor derartiger Geschäftsmodelle. Die Informationssuche mit mobilen Endgeräten ändert dabei die Bedingungen suchbezogener Werbung. Obwohl überwiegend in stationären Kontexten genutzt, findet sie nicht selten in sozialen Austauschsituationen statt und weist oftmals einen Lokationsbezug auf (Church & Oliver 2011). Weiterhin ist der Bildschirm von Smartphones erheblich kleiner als bei stationären oder portablen Rechnern, so dass weniger Platz für die Darstellung von Ergebnissen zur Verfügung steht und sich in Folge das Informationsverhalten noch stärker auf die höchstplatzierten Treffer fokussiert (Schwartz 2014).

Die Wahrnehmung und Akzeptanz mobiler Suchanzeigen durch die Nutzer ist bislang nur wenig erforscht. Gemäß eines Whitepapers von Marin Software (o.A. 2013) sind im Vergleich zur Desktopsuche mobil die Klickraten auf Ads höher, die Klickkosten geringer und Konversionsraten ebenfalls deutlich niedriger. Solche statistische Analysen weisen auf eine grundlegende Akzeptanz aber (derzeit) noch geringere Effektivität der mobilen Suchwortvermarktung hin. Nutzerorientierte Studien zum Gegenstandsbereich werden derzeit noch vermisst. Dies ist der Ausgangspunkt der vorliegenden Untersuchung. Ziel ist es in einem ersten Schritt, grundlegende Aspekte der Wahrnehmung und Akzeptanz mobiler Suchanzeigen auf Nutzerseite zu eruieren.

2. Forschungsdesign

Das Untersuchungsdesign orientiert sich grundlegend an methodischen Ansätzen, wie sie bei der nutzerzentrierten Untersuchung von Suchanzeigen und Display Advertising im Desktop-Bereich Verwendung finden – vgl. Malheiros et al. (2012), Owens et al. (2011), Buscher et al. (2010); Jansen & Resnick (2006). Um die Wahrnehmung und Akzeptanz von Suchanzeigen zu eruieren ist es sinnvoll, das tatsächliche Nutzerverhalten zu beobachten und die subjektive Einschätzung von Suchanzeigen zu erfragen. Die Beobachtung des Nutzerverhaltens liefert zunächst Hinweise zur Allokation der Aufmerksamkeit durch die Erfassung der Blickrichtung in Bezug auf die Ergebnisseite des verwendeten Suchdienstes, im Folgenden als SERP (Search Engine Result Page) bezeichnet. Das Selektionsverhalten der Einträge auf der SERP, im Speziellen die Klickrate auf die Werbeanzeigen, zeigt auf, inwiefern diese als relevant erachtet werden und sich tatsächlich effektiv für die Besuchergenerierung der Werbetreibenden auswirken. Die Befragung der Nutzer gibt einen Einblick dahingehend, ob diese die Werbeanzeigen bewusst (als solche) wahrnehmen und wie sie diese einschätzen. Fasst man diese Punkte zusammen, so lassen sich die folgenden drei Forschungsfragen ableiten.

1. Wie hoch ist die Aufmerksamkeit für Werbeanzeigen?
2. Wie effektiv sind Werbeanzeigen (für die Besuchergenerierung)?
3. Wie werden Werbeanzeigen (durch die Nutzer) eingeschätzt?

Zur Beantwortung der Forschungsfragen wurde das Untersuchungsdesign als Kombination von Nutzertests und -befragung wie folgt operationalisiert.

Jeder Nutzer führte zu vier vorgegeben Informationsbedürfnissen auf einem mobilen Endgerät eine Onlinesuche durch. Ein Google Nexus 4 wurde als einheitliches Testgerät spezifiziert, das zum Zeitpunkt der Untersuchung mit einer Größe von 4,7 Zoll und einer Auflösung von 1280x768 Pixel, eine „typische“ Gerätekonfiguration repräsentieren soll. Die Tests wurden mit dem Browser Firefox Mobile durchgeführt. Die Testszenarien wurden in Rückgriff auf die Selektion und Absichtsbeschreibung kommerzieller Anfragen gemäß Frisch (2013) festgelegt. Es handelt sich um informationelle oder transaktionale Suchanfragen (Broder 2002), die ein kommerziell verwertbares Interesse, z.B. eine Produktsuche, vermuten lassen und mindestens zwei Suchanzeigen aufweisen. Tabelle 1 zeigt die verwendeten Suchanfragen und die rekonstruierten Suchabsichten in der Übersicht.

Nr.	Suchbegriff	Suchabsicht
1	iphone 6 preis	Sie möchten allgemeine Informationen erhalten und interessieren sich vor allem für den Preis.
2	digitalkamera test	Sie möchten sich über die besten Kameras informieren und interessieren sich vor allem für Testberichte.
3	wii spiele	Sie möchten ein für Sie interessantes Spiel für Ihre Wii-Konsole finden.
4	abnehmen tipps	Sie möchten abnehmen und suchen nach Informationen, wie man am besten die Ernährung umstellt und welche Sportarten besonders beim Abnehmen helfen.

Tabelle 1: Testszenarien

Um ein möglichst einheitliches Testdesign zu gewährleisten, wurde ein Klick-Dummy generiert, in welchem die SERP zu den jeweiligen Suchbegriffen für alle Testpersonen einheitlich vorgegeben war. Die Aufgabe der Testpersonen bestand damit darin, mit der SERP zu interagieren, ggf. Treffer zu sichten und wieder auf die SERP zurück zu kehren. Weitergehende suchrelevante Interaktionen auf der SERP, wie die Reformulierung der Suchanfrage, waren nicht vorgesehen. Folgende Abbildung zeigt die Auswahl der 4 Szenarien und den unmittelbar sichtbaren Teil der SERPS zu den 4 Szenarien. Als Grundlage für den Vergleich des Nutzerverhaltens enthält die SERP zu Aufgabe 2 keine Werbeanzeigen als Ergebnisse, sondern nur organische Suchtreffer.

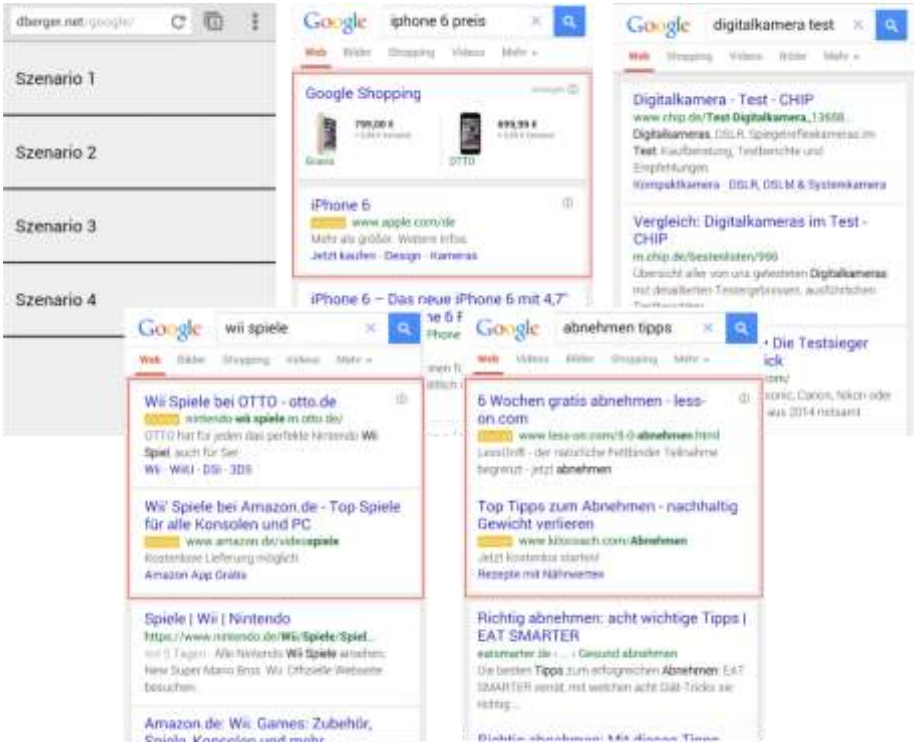


Abbildung 1: Klick-Dummy der SERPs mit markierten Werbeanzeigen

Der Testablauf wurde wie folgt spezifiziert. Nach einer kurzen Begrüßung und Einweisung in die Testumgebung führten die Probanden die Suchszenarien 1-4 durch. Hierbei wurde die Interaktion in Form von Videos protokolliert. Zudem wurden mittels eines Eye-Tracking-Gerätes (Tobii X2-60) die Fixationen der Nutzer relativ zu dem Display des Smartphones erfasst, welches dafür in einer Vorrichtung für das Eye-Tracking mit mobilen Endgeräten befestigt wird. Nach jeweils drei Minuten wurden die einzelnen Szenarien beendet. Nachdem alle vier Szenarien durchlaufen waren, wurde ein abschließendes Interview durchgeführt.

Die Tauglichkeit des Testdesigns wurde durch einen Pretest überprüft. Hier zeigte sich Verbesserungsbedarf bei den ursprünglichen Queries, die als zu unspezifisch erachtet wurden. So wurde z.B. die ursprüngliche Query „iphone“ zu „iphone preis“ modifiziert. Des Weiteren wurde eine zeitliche Beschränkung der Suchdauer zu einer Aufgabe auf drei Minuten eingeführt, da die Testperson sehr sorgfältig vorging und viel Zeit auf den Landingpages verbrachte, was für die Untersuchung weitgehend uninteressant war. Die Überbrückung der Wartezeitdauer an einer Bushaltestelle wurde hierzu als fiktives und möglichst glaubwürdiges Hintergrundszenario genutzt. Ebenso wurde ein ursprünglich vorgesehenes „Retrospective Think Aloud“ (RTA)

im Anschluss an die Untersuchung gestrichen, da die Probandin während des Pre-Tests hierbei ihre Auswahlentscheidung auf der SERP ausführlich begründete und das Vorgehen auf den Zielseiten elaborierte. Dies war für die Untersuchung nicht zielführend. Deshalb wurde stattdessen ein auf die Forschungsfragen ausgerichteter Fragebogen als Leitfaden für ein abschließendes Interview entwickelt.

In Anlehnung an die Ausführungen von Nielsen (2006) wurde eine Stichprobengröße von 20 Testpersonen angestrebt. Die Auswahl schränkt sich aufgrund des Themas auf Smartphone-Nutzer und aufgrund der Methode Eye-Tracking auf Personen mit gesunden Augen ein. Die Probanden wurden per persönlicher Ansprache und Email aus dem persönlichen und studienbezogenen Umfeld der Erstautorin rekrutiert. Die Testpersonen können in ihrer Gesamtheit als eher junge (23-34 Jahre Mittelwert, 26,25 Jahre), überwiegend weibliche (12 w, 8 m), hoch gebildete und im Umgang mit Smartphones routinierte Stichprobe bezeichnet werden. Von den Nutzern geben zwei an für die Informationssuche „immer“ das Smartphone zu nutzen, 14 antworteten mit „meistens“ und vier mit „manchmal“. Insofern können und sollen die Ergebnisse nicht unreflektiert auf die Gesamtheit aller Nutzer übertragen werden. Die Tests fanden zwischen dem 31. Oktober und dem 5. November 2014 statt. Insgesamt nahmen 21 Testpersonen an der Untersuchung teil. Ein Proband, der zur Studie erschien, benutzt privat kein Smartphone. Die Ergebnisse dieses Tests wurden deshalb nachträglich aus dem Sample eliminiert. Die Testdauer belief sich im Durchschnitt auf ca. 30 Minuten.

3. Analyse

Nachfolgend werden die Ergebnisse in Bezug auf die Forschungsfragen 1-3 dargelegt und anschließend zusammengeführt.

3.1 Wie hoch ist die Aufmerksamkeit für Werbeanzeigen?

Grundlage der Analyse zur ersten Forschungsfrage bilden die durch Eye-Tracking erfassten Fixationen der Nutzer auf der SERP zur jeweiligen Suchaufgabe. Nachfolgende Abbildung zeigt die kumulierte Blickverteilung vor dem ersten Scrollen für die vier Suchaufgaben.

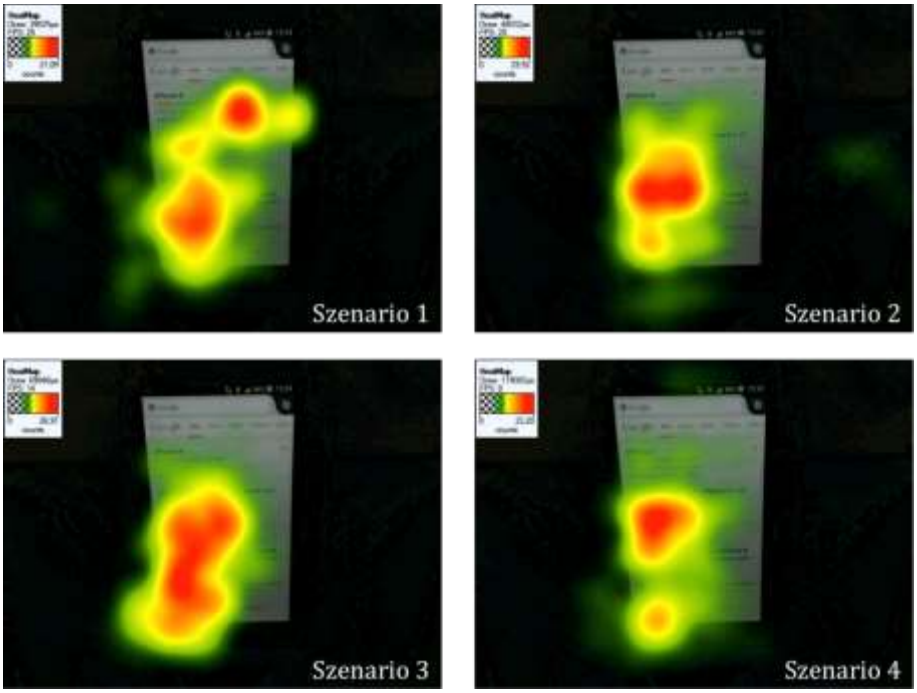


Abbildung 2: Blickverteilung vor dem ersten Scrollen.

Die Werbeanzeigen nehmen im ersten Szenario 57% Fläche des unmittelbar, (ohne zu scrollen) sichtbaren Bereichs der SERP ein. Für Szenario 3 und 4 beträgt der Anteil der Werbung 60%. Mit Ausnahme des ersten Szenarios liegt die hauptsächliche Wahrnehmung in der Mitte des Bildschirmes. Dies gilt auch für Szenario 2, in dem an dieser Stelle keine Werbeanzeige eingeblendet ist. Da alle Probanden die Testszenarien in der Reihenfolge Szenario 1 bis Szenario 4 durchliefen ist es möglich, dass die Probanden lernen, dass das oberste Ergebnis des ersten Szenarios einen Werbeeintrag darstellt und dies auch bei den Folgeszenarien erwarten. Eine weitere mögliche Erklärung besteht darin, dass der erste Werbeeintrag in Szenario 1 durch die grafischen Displays mehr Aufmerksamkeit bekommt. Diese Vermutung geht mit den Ergebnissen von Malheiros et al. (2012) konform, die eine höhere Aufmerksamkeit für targeted rich media ads konstatieren.

Insgesamt lässt sich feststellen, dass Werbung, die über den organischen Suchergebnissen eingeblendet wird, sensorisch wahrgenommen wurde. Der Schwerpunkt der Wahrnehmung, gemessen an der Anzahl der Fixationen, liegt dennoch auf den organischen Suchtreffern, die zunächst sichtbar sind. Werbeeinträge, die am Ende der SERP, d.h. unterhalb der organischen Treffer eingeblendet werden, wurden nur teilweise bemerkt. In drei der Szenarien scrollte nur knapp die Hälfte der Probanden bis ans Ende der Ergebnisseite

und konnte damit auch die unten positionierten Anzeigen in den Szenarien 1 und 4 wahrnehmen. In Szenario 3 erreichte lediglich eine Testperson die Werbung am Ende der SERP.

Zur methodischen Einschätzung lässt sich festhalten, dass die Interaktion auf der SERP individuell sehr unterschiedlich ausfällt. So variierte die Zeitdauer bis zum ersten Scrollen von 1-2 Sekunden als kürzeste gemessene Zeitspanne, bis zu 7-14 Sekunden als längste gemessene Zeitspanne in den Szenarien. Der Mittelwert über alle Szenarien beträgt 4,76 Sekunden (SA=3,25 Sekunden). Dasselbe gilt für die Dauer bis zur ersten Trefferauswahl. Die gemessenen Werte reichen von 2-5 Sekunden als kürzeste und 32-44 Sekunden als längste gemessene Zeitspanne. Der Mittelwert der Dauer bis zur ersten Selektion beträgt 16,7 Sekunden (SA=9,7). Während die Zeit bis zum Scrollen personenabhängig zu sein scheint, trifft das für die Dauer bis zum ersten Klick eher nicht zu. Mehr wahrgenommene Treffer führen demnach nicht notwendigerweise zu einer längeren Dauer für die Auswahl eines Treffers, was bei einer Wahrnehmungsanalyse berücksichtigt werden muss. Insgesamt gilt damit, dass die wahrnehmungsbezogene Analyse nicht nur tentativ im Sinne einer ersten methodischen Näherung angelegt ist, sondern auch die Ergebnisse vorsichtig interpretiert werden müssen.

3.2 Wie effektiv sind Werbeanzeigen?

Die Effektivität der Werbeanzeigen wird an der tatsächlichen Selektion derselben durch die Nutzer festgemacht. Insgesamt klickten die Nutzer 176 Mal auf Suchergebnisse. Während des gesamten Testes wurde bei den drei Szenarien, die Werbung enthalten, von allen 20 Testpersonen nur neunmal Werbeanzeigen ausgewählt. Dies entspricht einem prozentualen Anteil von 5,11% Klicks bei diesen drei Aufgaben (8,7% Szenario 1, 9,8% Szenario 3, 2,2% Szenario 4). Trotz der hohen Sichtbarkeit der Top-Werbeeinträge zeigt sich also eine klare Präferenz für organische Ergebnisse. Diese wird noch dadurch verdeutlicht, dass in lediglich zwei Fällen ein Werbeeintrag als erster Treffer selektiert wurde. In den anderen Fällen wurden Werbeeinträge erst ausgewählt, nachdem organische Treffer selektiert wurden. Abbildung 3 veranschaulicht die Klickverteilung in den einzelnen Testszenarien.

Man sieht deutlich eine Präferenz für die Top-Positionen, sofern es sich bei diesen Treffern nicht um Werbung handelt. Immerhin erzielen alle drei Such-Szenarien, in denen Werbung enthalten ist, Klicks auf die Werbeeinträge zu Beginn der Seite. Nur bei einer der drei Aufgaben zeigen sich auch Klicks auf einen der unteren Werbeeinträge. Zur Verdeutlichung aggregiert Abbildung 4 die Klickrate der Szenarien, die Werbung enthielten.

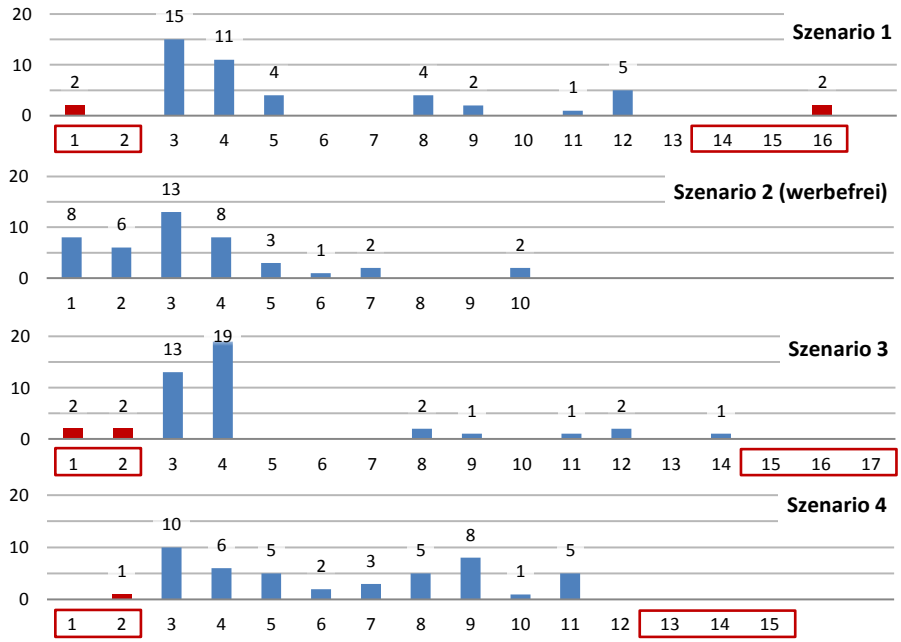


Abbildung 3: Klickverteilung nach Trefferposition für die vier Testszenarien, Werbeeinträge rot markiert (Szenario 1, 3 und 4)

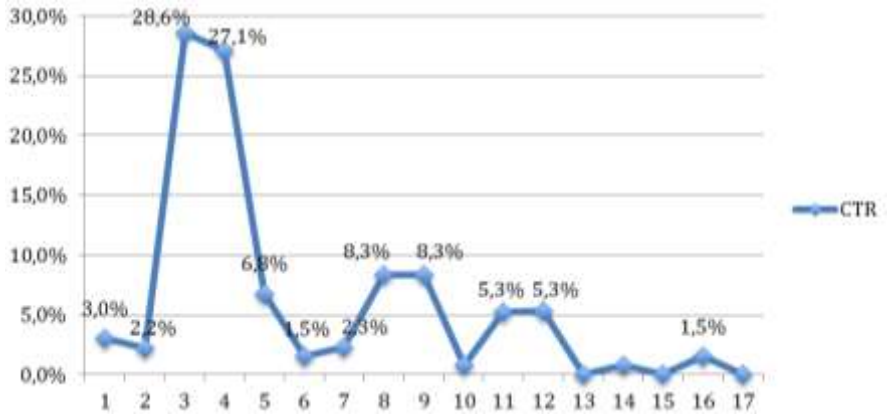


Abbildung 4: Aggregierte Klickverteilung (Click Through Rate, CTR)

Schließlich bleibt festzuhalten, dass nur 5 von 20 Testpersonen überhaupt Werbeeinträge selektieren. Vor diesem Hintergrund deuten die Ergebnisse in hohem Maße darauf hin, dass die Nutzer bei der Suche auf dem Smartphone Paid Listings in hohem Maße vermeiden. Gerade für Nutzer, die keine Werbung selektieren bleibt zu konstatieren, dass die Google Suche in den Fällen, in denen Werbung eingebunden wird, den Nutzern, ohne scrollen zu müssen

(above the fold), nur ein organisches Ergebnis anbietet. Das ist deutlich weniger als im Vergleich zur Desktop-Suche.

3.3 Wie werden Werbeanzeigen eingeschätzt?

In dem abschließenden Interview wurden die Probanden zu ihrer subjektiven Einschätzung befragt. Zunächst wurde gefragt, ob die Testpersonen die eingeblendete Werbung auch bewusst wahrgenommen haben. Hierbei antworteten 16 (80%) mit ja. Diese wurden befragt, ob sie die Werbung als störend empfanden. Dies bestätigte wiederum nur eine Person. Dies deutet darauf hin, dass Nutzer suchbezogene Werbung bzw. deren geschäftsbezogene Notwendigkeit akzeptieren. Dennoch haben 12 dieser 16 Personen angegeben, die Werbung bewusst gemieden zu haben. Die Angaben waren hierbei weitgehend kohärent zum tatsächlichen Verhalten. Die Testpersonen, die angaben, auf Werbung bewusst zu verzichten, klickten auch tatsächlich keine Werbung an. Lediglich eine Person wählte Anzeigen in den Szenarien 1, 3 und 4, obwohl sie angab, bei Google-Suchen erst ab dem dritten Ergebnis aufmerksam zu lesen. Von den vier Testpersonen, die angaben, Werbung nicht bewusst zu meiden, klickten drei Probanden (75%) auf mindestens eine Anzeige während des Tests. Die Befragung unterstützt auch die These des „Lerneffekts“ für die werbefreie zweite Suchaufgabe. 13 Testpersonen fiel nicht auf, dass eines der vier Suchszenarien keine Werbung enthielt. In diesem Fall kann man mutmaßen, dass die Erwartung von Paid Listings die Selektionschancen der organischen Treffer 1 und 2 bei der werbefreien Suchaufgabe reduzierten. Die weiteren Fragen fokussierten sich auf die generelle Einstellung gegenüber Werbung. 15 (75%) Testpersonen gaben an, dass sie generell auf Werbung klicken würden, wenn bestimmte Gegebenheiten vorhanden sind. Das am häufigsten genannte Kriterium ist Seriosität. Viele Testpersonen nannten dabei als Positivbeispiel den Online-Shop Amazon.

4. Zusammenfassung und Diskussion

Führt man die einzelnen Teilergebnisse der Untersuchung zusammen, so lässt sich konstatieren, dass suchbezogene Werbung, sofern diese am Anfang der SERP eingeblendet wird, wahrgenommen wird. Der Schwerpunkt der Aufmerksamkeit liegt dennoch auf den organischen Suchtreffern, so dass sich auch für den mobilen Bereich eine Vermeidung von Werbung bei der Internetsuche ergibt. Dieses Bild verstärkt sich, wenn man das Verhalten der Probanden analysiert. Trotz der Tatsache, dass die Werbeeinträge vor dem ersten Scrollen einen Großteil der sichtbaren Fläche einnehmen und damit quasi kaum zu vermeiden sind, werden Paid Listings kaum geklickt. Zwar erzielen alle drei Such-Szenarien in denen Werbung enthalten ist, Klicks auf die Wer-

beeinträge zu Beginn der Seite. Auch die Klickrate entspricht über alle Aufgaben hinweg in etwa dem von Marin Software beschrieben erhöhten Niveau[22] im Vergleich zur Desktopsuche (o.A. 2013). Dennoch vermeiden 75% Nutzer die Werbung bewusst. Von den 20 Probanden wird bei den insgesamt 60 durchlaufenen Suchaufgaben, die Werbung enthalten, nur zweimal als erster Treffer ein Werbeeintrag selektiert. In den anderen Fällen wird zunächst ein organisches Ergebnis präferiert. Die Effektivität von mobilen Suchanzeigen bei der Generierung von Besuchern ist damit im Verhältnis zu ihrer prominenten Platzierung eher gering. Dies gilt gerade auch dann, wenn man in Betracht zieht, dass die Paid Listings die organischen Suchergebnisse auf dem Smartphone in deutlich höherem Maße aus dem unmittelbar sichtbaren Bereich verdrängen, als dies bei der Suche auf dem Desktop der Fall ist. Die Befragung der Nutzer induziert, dass Werbung zwar wahrgenommen und quasi auch erwartet, aber oft auch bewusst vermieden wird. Nichtsdestotrotz wird die Werbung akzeptiert, d.h. als nicht grundsätzlich störend empfunden. Insgesamt bestätigen die vorliegenden Ergebnisse im großen Ganzen die Ergebnisse, die zur Wahrnehmung suchbezogener Werbung auf dem Desktop vorliegen, vgl. Malheiros et al. (2012), Owens et al. (2011), Buscher et al. (2010); Jansen & Resnick (2006). Auch die von o.A. (2013) argumentierten hohen Klickraten auf mobilen Endgeräten können hier nachvollzogen werden.

Aus methodischer Perspektive ist die vorliegende Untersuchung als eine erste Exploration im Themenfeld einzuordnen. In weiteren Untersuchungen ließen sich weitere Faktoren, wie die Relevanz der Werbeeinträge mit einbeziehen bzw. verfeinerte Analysen in Bezug auf das Verhalten unterschiedlicher Nutzertypen (z.B. Ad-Klicker vs. Ad-Non-Klicker) durchführen.

5. Literaturverzeichnis

Broder, A. (2002). A taxonomy of web search. In ACM Sigir forum (Vol. 36, No. 2), S. 3-10.

Buscher, G.; Dumais, S. T. & Cutrell, E. (2010). The good, the bad, and the random: an eye-tracking study of ad quality in web search. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, USA, S. 42-49.

[22] Auch wenn man die Verteilung der Klickrate in der Untersuchung aufgrund der Eigenschaften des Testdesign nicht unreflektiert 1:1 auf die CTR-Messung in dem Whitepapers von Marin Software (o.A. 2013) übertragen kann.

Church, K. & Oliver, N. (2011). Understanding mobile web and mobile search use in today's dynamic mobile landscape, in Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services, ACM, New York, NY, USA, S. 67-76.

Dischler, J. (2015). Building for the next moment, Inside AdWords, URL http://adwords.blogspot.de/2015/05/building-for-next-moment.html (letzter Zugriff 11.05.2015)

Frisch, S. (2013). Mobile Search Advertising – Vergleich der Webseiten von organischen und bezahlten Suchtreffern, Masterarbeit Universität Hildesheim.

Google (2015). Q1 2015. Quarterly Earnings Summary. URL https://investor.google.com/pdf/2015Q1_google_earnings_slides.pdf (letzter Zugriff 08.05.2015)

Jansen, B. J., & Resnick, M. (2006). An examination of searcher's perceptions of nonsponsored and sponsored links during ecommerce Web searching. Journal of the American Society for Information Science and Technology, 57(14), S. 1949-1961.

Malheiros, M., Jennett C., Patel, S., Brostoff, S., Sasse, M.A. (2012). Too Close for Comfort: A Study of the Effectiveness and Acceptabilitiy of Rich-Media Personalized Advertising. In: ‘Proceedings of SIGCHI Conference on Human Factors in Computing Systems', ACM, New York, NY, USA, S. 579-588.

Nielsen, J. (2006). Quantitative Studies: How Many Users to Test? URL http://www.nngroup.com/articles/quantitative-studies-how-many-users/

o.A. (2013). Mobile Search Advertising around the Globe. How Smartphones & Tablets Are Changing Paid Search. 2013 Annual Report. Marin Software Whitepaper, URL http://www.iabaustralia.com.au/uploads/uploads/2013-10/1382054400_daf59f00d5fda6ce1b6f0cbae937f0ec.pdf (letzter Zugriff 10.06.2015)

Owens, J. W., Chaparro, B. S., & Palmer, E. M. (2011). Text advertising blindness: the new banner blindness?. Journal of Usability Studies, 6(3), S. 172-197.

Schwartz, B. (2014). Mobile Search Ranking Study: Rank Number One Or Not Rank At AllRanking in position one in mobile is way more important than on desktop search. URL http://searchengineland.com/mobile-search-ranking-study-rank-number-one-rank-206510 (letzter Zugriff 10.06.2015)

PRAXISTRACK

Factors influencing the adoption and acceptance of an Enterprise 2.0 tool for knowledge exchange

Examining system-, organization- and user-related predictors

Saskia Untiet-Kepp, Melanie Mönch

Hellmann Worldwide Logistics GmbH & Co. KG
Elbestraße 3
Osnabrück
saskia.untiet-kepp@de.hellmann.net

Abstract

While the external use of social media tools by businesses has already been widely studied, the use of internal social media for the purpose of facilitating knowledge exchange within companies, which is often referred to as Enterprise 2.0, has not. Previous research shows the benefits, Enterprise 2.0 could bring but also that these goals are only rarely achieved, because of low adoption rates. The aim of this paper is to investigate the factors influencing the adoption of Enterprise 2.0 tools by proposing a holistic view that takes into account a diverse set of system-related, organization-related as well as user-related predictors, which influence the adoption of Enterprise 2.0 tools and are highly interdependent.

1. Introduction

As much as the rise of social media has changed computer-mediated communication in our private life (cp. Richter et al. 2011; Razmerita et al. 2014), it also changes how we do business (Turban et al. 2011). By the use of social media technologies, businesses can improve external as well as internal communication in order to strengthen customer relationships and raise productivity (Turban et al. 2011; McKinsey Global Institute 2012). While the external use of social media tools has already been widely studied, the use of internal social media for the purpose of facilitating knowledge exchange, which is often referred to as Enterprise 2.0, has not (Zhang et al. 2009). Previous research shows the benefits, Enterprise 2.0 could bring but also that these goals are only rarely achieved, because of low adoption rates (Wang et al. 2014). The aim of this paper is to investigate the factors influencing the

adoption of Enterprise 2.0 tools. Previous research of the adoption and its influencing factors is still scarce and mostly concentrates on single or only a few dimensions (Turban et al. 2011; Richter et al. 2011). This paper is part of a study carried out at a global logistics company in 2015 and proposes a holistic view taking into account a diverse set of predictors that are system-related, organization-related as well as user-related

2. Explanation of terms

Before the main study method and results are presented, a few terms need to be explained in the context of this study.

2.1 Enterprise 2.0

Richter et al. (2011, p. 91) define Enterprise 2.0 in the following way: "Enterprise 2.0 refers to the phenomenon of a new participatory corporate culture (with regard to communication and information sharing), which is based on the application of various types of social software technologies." The purpose of Enterprise 2.0 can be described as exchanging knowledge "through social interaction and collaboration among employees mediated by social media" (Razmerita et al. 2014, p. 79).

2.2 Knowledge Exchange

Knowledge exchange describes the reciprocal processes of sharing and seeking knowledge (Wang & Noe 2010). Sharing knowledge means to provide knowledge, whereas seeking knowledge includes searching and collecting knowledge.

2.3 Adoption and Acceptance

Adoption and Acceptance of a technology can be described as two distinct processes within individuals. The first one is adoption, which consists of the following steps according to Rogers (1995, p. 21) after Frambach & Schillewaert (2002, p.164):

- First knowledge of an innovation
- Forming an attitude towards the innovation
- Decision to adopt or reject
- Implementation of the new idea

- Confirmation of the decision

At the end of the adoption process the individual either has the intention to use the new technology (adoption) or not (rejection).

Acceptance on the other hand starts at the point where the intention has already been formed and is therefore taking place post-adoption. Acceptance is described by Venkatesh (2003) as a process that consists of the following steps:

- Intention to use the technology
- Using the technology
- Forming an intention whether to use the technology again

Since the study was carried out in a setting where active users as well as non-users are present, the study investigates both adoption and acceptance.

3. The Research Question and Model

The research question addressed in the presented study was coined as follows: What are the factors influencing the acceptance and adoption of an Enterprise 2.0 tool for knowledge exchange?

The research model used is based on the Unified Theory of Acceptance and Use of Technology (UTAUT) (Venkatesh et al. 2003) that has been refined and extended for this study by variables that have been found to predict and describe knowledge exchange in previous studies.

Behavioral intention to use Enterprise 2.0 tools for knowledge exchange is the dependent variable in this study. The goal of this research paper is to find those independent variables that predict behavioral intentions of actual users as well as non-users.

The following variables were included in the study:

3.1 System-Related Predictors

- Performance Expectancy:
  "degree to which an individual believes that using the system will help him or her to attain gains in job performance" (Venkatesh et al., 2003, p. 447)
- Effort expectancy:
  "degree of ease associated with the use of the system" (Venkatesh et al., 2003, p. 450).

3.2 Organization-Related Predictors

The predictor social influence was distinguished into the following two variables:

- Subjective norm
  "degree to which an individual perceives that important others believe he or she should use the new system." (Venkatesh et al. 2003, p. 451)
- Perceived network externality
  "users' perceptions of whether an information technology has attracted a sufficient number of users to indicate that [a] critical mass has been reached". (Wang et al. 2014, p. 1054)

Further predictors are organizationally provided technical support and organizational climate. The first incorporates the personal support an organization offers through for example trainings, instructions or a contact person, who has the abilities to consult (Venkatesh et al. 2003). The latter is defined as "a contextual situation at a point in time and its link to the thoughts, feelings, and behaviors of organizational members." (Bock et al. 2005, p. 89)

3.3 User-Related Predictors

As user-related predictors the survey contained trust in colleagues, individual differences and demographic data. Trust in colleagues is defined as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party." (Mayer, Davis, & Schoorman 1995, p. 712)

Individual differences are divided into the following two:

- Innovativeness
  "tendency of a person to adopt an innovation within a product class, independently of the communicated experience of others." (Frambach & Schillewaert 2002)
- Organizational identification
  "degree to which a member defines him- or herself by the same attributes that he or she believes define the organization." (Dutton, Dukerich, & Harquail 1994, p. 239)

As demographic data gender, age, work experience, education, job type and usage frequency were used.

4. Method

The study was carried out as a survey within a global logistics company. The results presented here were gained from 399 survey participants from Germany.

Since behavioral intention is the only dependent variable under investigation, the results were analyzed by use of a hierarchical multiple regression analysis. The observations were divided into actual users and non-users. According to the definitions given above, we define the behavioral intention to exchange knowledge for actual users as Acceptance of Enterprise 2.0 tools and behavioral intention to exchange knowledge for non-users as Adoption of Enterprise 2.0 tools.

For this analysis the variables were organized into three sets:

1st Set:

- Performance Expectancy
- Effort expectancy
- Subjective Norm
- Perceived Network Externality
- Provided technical support

2nd Set:

- Organizational climate
- Trust in colleagues
- Innovativeness
- Organizational identification

3rd Set:

- Gender
- Job Type
- Education
- Usage Frequency
- Work Experience
- Age

The following section describes the results of the hierarchical multiple regression analysis.

5. Results

The Behavioral intention to exchange knowledge was distinguished between the intention to give knowledge and the intention to get knowledge. For actual users, the results for these both aspects of knowledge exchange showed significant differences, so that they will be described here separately. For the non-users, an exploratory factor analysis showed that the behavioral intention to give and get knowledge could not be analyzed separately. Hence, knowledge exchange was viewed as one single concept for non-users.

The results depicted in Figure 2 show that performance expectancy, subjective norm and usage frequency positively influence the behavioral intention to get knowledge via the enterprise 2.0 system for actual users. The influence for all of these variables, except the usage frequency, is positive.

The results in Figure 3 show that performance expectancy, effort expectancy, perceived network externality and usage frequency influence the behavioral intention to give knowledge via the enterprise 2.0 system for actual users. Performance expectancy is positively influencing giving knowledge, whereas effort expectancy, perceived network externality and usage frequency have a negative impact on the behavioral intention.

Models	B	SE B	β	R^2 (ΔR^2)
Step 1				
Constant	1.67	.33		
Performance expectancy	.32	.08	.33***	.22 (.19)
Effort expectancy	-.01	.07	-.02	
Subjective Norm	.19	.08	.20*	
Perceived Network externality	.07	.07	.08	
Provided technical support	-.10	.07	-.10	
Step 2				
Constant	1.64	.53		
Performance expectancy	.32	.08	.33***	
Effort expectancy	-.03	.07	-.03	
Subjective Norm	.16	.08	.17*	
Perceived Network externality	.03	.08	.03	
Provided technical support	-.10	.07	-.11	
Organizational climate	.13	.09	.13	.24 (.20)
Trust in colleagues	.03	.10	.02	
Innovativeness	.08	.09	.07	
Organizational Identification	-.14	.08	-.13	
Step 3				
Constant	-8.32	14.83		
Performance expectancy	.23	.08	.23**	
Effort expectancy	.03	.07	.03	
Subjective Norm	.11	.08	.12	
Perceived Network externality	.05	.08	.05	
Provided technical support	-.13	.07	-.13	
Organizational climate	.12	.08	.12	
Trust in colleagues	.09	.10	.07	
Innovativeness	.03	.08	.03	
Organizational Identification	-.13	.08	-.12	
Gender (Male)	.00	.12	.00	.32 (.25)
Job Position	-.16	.13	-.10	
Education	-.01	.04	-.02	
Usage Frequency	-.27	.07	-.30***	
Work experience	.00	.01	.01	
Age	.01	.01	.08	

*p < 0.05. **p < 0.01. ***p < 0.001.

Figure 2: Results of the hierarchical multiple regression for getting knowledge for actual users

Models	B	SE B	β	R^2 (ΔR^2)
Step 1				
Constant	1.69	.49		
Performance expectancy	.50	.12	.35***	.18 (.15)
Effort expectancy	-.27	.10	-.21**	
Subjective Norm	.20	.12	.15	
Perceived Network externality	-.27	.11	-.21*	
Provided technical support	.12	.10	.09	
Step 2				
Constant	2.45	.79		
Performance expectancy	.51	.12	.36***	
Effort expectancy	-.27	.10	-.22**	
Subjective Norm	.21	.12	.15	
Perceived Network externality	-.30	.12	-.23**	
Provided technical support	.13	.12	.09	
Organizational climate	.10	.13	.07	.19 (.14)
Trust in colleagues	-.21	.15	-.11	
Innovativeness	-.00	.13	-.00	
Organizational Identification	-.11	.12	-.07	
Step 3				
Constant	28.19	21.97		
Performance expectancy	.38	.12	.27**	
Effort expectancy	-.19	.10	-.15	
Subjective Norm	.13	.12	.09	
Perceived Network externality	-.32	.11	-.24**	
Provided technical support	.07	.10	.06	
Organizational climate	.08	.12	.05	
Trust in colleagues	-.10	.15	-.05	
Innovativeness	-.04	.13	-.02	
Organizational Identification	-.06	.12	-.04	
Gender (Male)	-.17	.18	-.07	.30 (.23)
Job Position	-.00	.19	-.00	
Education	-.02	.06	-.03	
Usage Frequency	-.44	.10	-.33***	
Work experience	-.01	.01	-.11	
Age	-.01	.01	-.12	

*p < 0.05. **p < 0.01. ***p < 0.001.

Figure 3: Results of the hierarchical multiple regression for giving knowledge for actual users

Models	B	SE B	β	R^2 (ΔR^2)
Step 1				
Constant	.31	.23		
Performance expectancy	.44	.06	.45***	.37 (.36)
Effort expectancy	.04	.07	.03	
Subjective Norm	.13	.06	.13*	
Perceived Network externality	.00	.07	.00	
Provided technical support	.20	.07	.18**	
Step 2				
Constant	-.22	.34		
Performance expectancy	.41	.06	.42***	
Effort expectancy	.01	.06	.01	
Subjective Norm	.12	.06	.12	
Perceived Network externality	-.03	.07	-.02	
Provided technical support	.16	.07	.15*	
Organizational climate	.01	.07	.01	.39 (.37)
Trust in colleagues	.17	.10	.11	
Innovativeness	.04	.06	.04	
Organizational Identification	.06	.07	.06	
Step 3				
Constant	23.23	10.16		
Performance expectancy	.39	.06	.40***	
Effort expectancy	.05	.07	.05	
Subjective Norm	.10	.07	.10	
Perceived Network externality	-.01	.07	-.01	
Provided technical support	.15	.07	.14*	
Organizational climate	.02	.07	.02	
Trust in colleagues	.18	.10	.12	
Innovativeness	.01	.06	.01	
Organizational Identification	.06	.07	.06	
Gender (Male)	.18	.11	.10	.42 (.38)
Job Position	-.03	.12	-.02	
Education	.01	.03	.02	
Work experience	-.01	.01	-.07	
Age	-.01	.01	-.16*	

*p < 0.05. **p < 0.01. ***p < 0.001.

Figure 4: Results for the hierarchical multiple regression analysis for knowledge exchange for non-users

The results in Figure 4 show that for non-users the following variables predict behavioral intent to exchange knowledge via the Enterprise 2.0 platform: Performance expectancy, subjective norm, provided technical support and age. All of these, except age have a positive impact on behavioral intention.

6. Discussion

With these results, the company gains exciting new input on what is really important for their employees in order to be willing to exchange knowledge via the collaborative platform. Moreover they can better derive necessary measures to drive adoption further. The most central insights are that users need to be aware of their potential performance gains by using the Enterprise 2.0 tool and of the training and support that is provided. Another conclusion to be drawn could be to further investigate the reasons behind the results. Summarizing the study, the results provide a comprehensive overview over the status of the Enterprise 2.0 endeavor and also give exciting insights into the current corporate climate.

7. References/Literaturverzeichnis

Bock, G.-W., Zmud, R. W., Kim, Y.-G., & Lee, J.-N. (2005). Behavioral Intention Formation in Knowledge Sharing: Examining the Roles of Extrinsic Motivators, Social-Psychological Forces, and Organizational Climate. *Management Information Systems Quarterly, 29(1)*, pp. 87-111.

Dutton, J., Dukerich, J., & Harquail, C. (1994). Organizational images and member identification. *Administrative Science Quarterly, 39(2)*, pp. 239–263.

Frambach, R. T., & Schillewaert, N. (2002). Organizational innovation adoption: a multi-level framework of determinants and opportunities for future research. *Journal of Business Research, 55(2)*, pp. 163-176.

Mayer, R., Davis, J., & Schoorman, F. (1995). An integrative model of organizational trust. *Academy of Management Review, 20(3)*, pp. 709-734.

McKinsey Global Institute. (2012). *The social economy: Unlocking value and productivity through social technologies.* Retrieved 16.11.2014, from http://www.mckinsey.com/insights/high_tech_telecoms_internet/the_social_economy

Razmerita, L., Kirchner, K., & Nabeth, T. (2014). Social Media in Organizations: Leveraging Personal and Collective Knowledge Processes. *Journal of Organizational Computing and Electronic Commerce, 24(1)* , pp. 74-93, doi: 10.1080/10919392.2014.866504.

Richter, D., Riemer, K., & vom Brocke, J. (2011). Internet Social Networking. Research State of the Art and Implications for Enterprise 2.0. *Business & Information Systems Engineering, 2*, pp. 89-101, doi: 10.1007/s12599-011-0151-y.

Turban, E., Bolloju, N., & Liang, T.-P. (2011). Enterprise Social Networking: Opportunities, Adoption, and Risk Mitigation. *Journal of Organizational Computing and Electronic Commerce, 21(3)*, pp. 202-220, doi: 10.1080/10919392.2011.590109.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *Management Information Systems Quarterly, 27(3)*, pp. 425-478.

Wang, S., & Noe, R. A. (2010). Knowledge sharing: A review and directions for future research. *Human Resource Management Review, 20*, pp. 115–131.

Wang, T., Jung, C.-H., Kang, M.-H., & Chung, Y.-S. (2014). Exploring determinants of adoption intentions towards Enterprise 2.0 applications: an empirical study. *Behaviour & Information Technology, 33(10)*, S. 1048-1064, doi: 10.1080/0144929X.2013.781221.

Zhang, A. M., Zhu, Y., & Hildebrandt, H. (2009). Enterprise Networking Web Sites and Organizational Communication in Australia. *Focus on Business Practices*, pp. 114-119, doi: 10.1177/1080569908330381.

Proceedings des
HiER – Hildesheimer Evaluierungs- und Retrievalworkshop 2015
Universitätsverlag Hildesheim
Universitätsplatz 1
31141 Hildesheim
verlag@uni-hildesheim.de
ISBN (Open Access) 978-3-934105-59-1
ISBN-A (Open Access) 10.978.3934105/591
Hildesheim 2015